

11 more models

Model: Think of three levels of modeling.

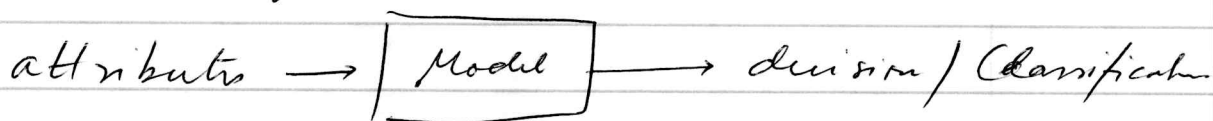
eg: ① Regression

② What are the factors/param?

③ The equation used ~~for~~ ^{to} actual prediction.

Classification

put things / sort things / differentiate
Binary: Yes/no ... examples (email)



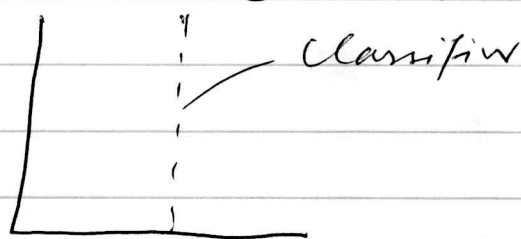
Cost of Classification: "Ehh" is not ok.

Edible / ~~Prison~~ Prisoner.

Cost is the big factor.

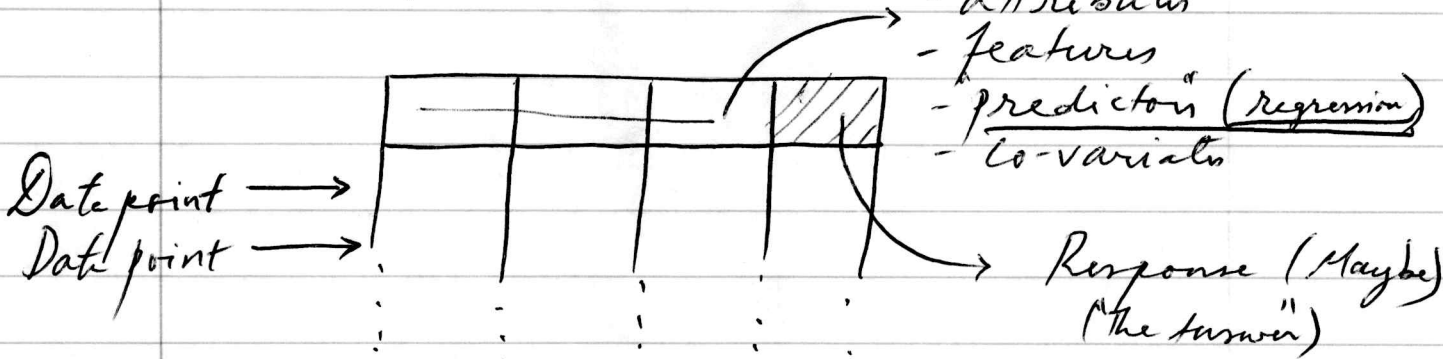
If cost is higher, move classifier away from it.

BE CONSERVATIVE!



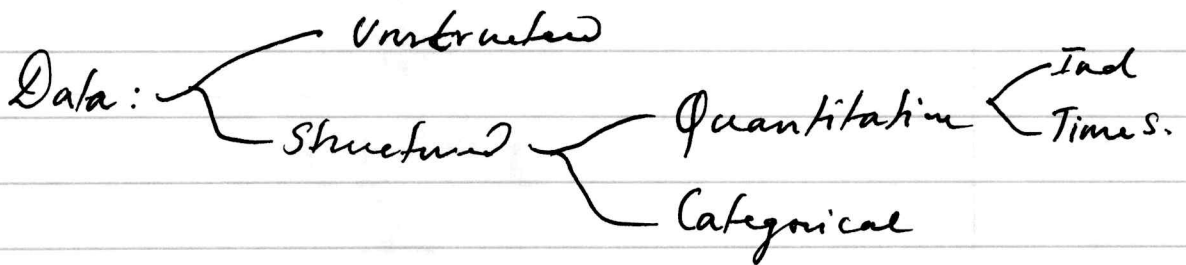
← uses only one attribute / predictor.

The Shape of Data



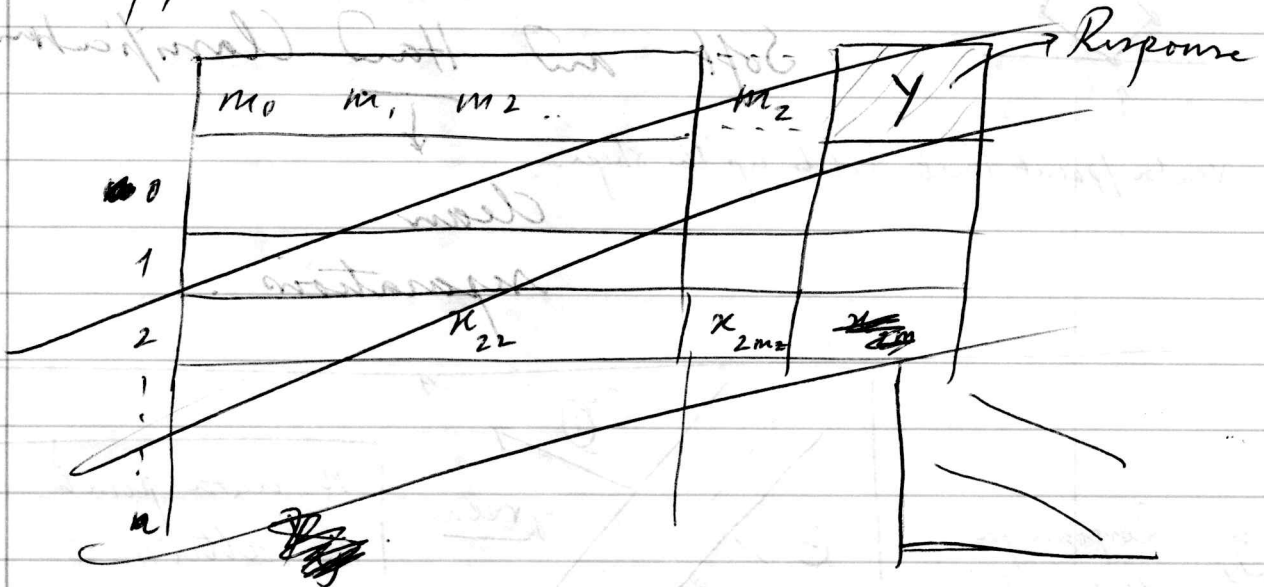
- TYPES**
- Structured
 - Quantitative ^{"number with meaning"} (age, score)
 - Categorical (incl. binary) (M/F, color)
 - ^{"number w/o meaning" eg. zip code.}
 - Numbers are not always Quantitative. (eg. ZIP code)
 - Independent data points are UNRELATED (Then there's time series data.)

Unstructured: like text



$$(a_1 x_{1j} + a_2 x_{2j} \dots a_m x_{mj}) y_j \geq 1 \quad \text{Classifier!}$$

Support Vector Machines



	0	1	2	...	(m)	RESPONSE
0	x_{00}				$x_{0,m}$	y_0
1						y_1
2						y_2
3						y_3
...						y_i
(n)						y_n

$$a_1 x_1 + a_2 x_2 + a_0 = 0$$

$$a_1 x_1 + a_2 x_2 + a_0 \geq 1$$

$$\sum_{i=1}^m a_i x_i + a_0$$

$$y_i = \begin{cases} 1 & \text{if ok} \\ -1 & \text{if not} \end{cases}$$

$$y = \begin{cases} w^T x + b = 1 \\ w^T x + b = -1 \end{cases}$$

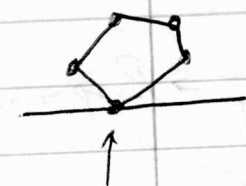
credit income paid

$$a_1 (\quad)$$

$$a_1 x_{1j} + a_2 x_{2j} + a_0 \geq 1 \quad \text{Gr}$$

$$a_1 x_{1j} + a_2 x_{2j} + a_0 \leq -1 \quad \text{Red}$$

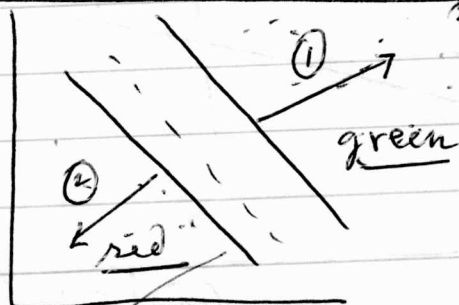
Soft and Hard Classification



vector/point that holds up the slope.

clean separations.

y_j : Response for datapoint j
1 if green
-1 if red



n data points
 m attributes

x_{ij} → datapoint $j \in [0, n]$
→ attribute $i \in [0, m]$

line so sum of coeffs. "How much does each coeff. contribute?"

one line is: $a_0 + a_1 x_1 + a_2 x_2 + \dots + a_m x_m = 0$
attribute! $a_0 + \sum_{i=1}^m a_i x_i = 0$

① $[a_0 + a_1 x_{1j} + a_2 x_{2j} + \dots + a_m x_{mj}] \geq m \rightarrow$ some threshold use 1 for scaling.
② $[\quad] \leq -m$ (or -1)

COMBINE: all points: $y_j [a_0 + a_1 x_{1j} + a_2 x_{2j} + \dots + a_m x_{mj}] \geq 1$
↓
use as binary variable!

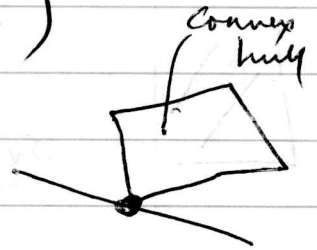
CONTD

SVM : $\sum a_i^2$: ~~Minimize this to~~

↳ min this to max margin!

$$\sum \max(0, 1 - \sum (a_i x_{ij} + a_0) y_j)$$

↳ this is all errors!



So: Minimize $\sum a_i^2$

Subject to $\sum (a_i x_{ij} + a_0) y_j \geq 1$

reduce errors.

(OR)

$$\text{Min} \left[\sum \max(0, 1 - \sum (a_i x_{ij} + a_0) y_j) + \lambda \sum a_i^2 \right]$$

enlarge margin

→ Need to scale data sometimes.

→ Can use regression (eq.) and do away w/ classification altogether.

separate this too!

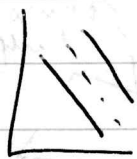
$$50 \times \sum_{i=1}^{25} + 25 \sum_{i=26}^{70} + 1,870 \sum_{i=71}^{400}$$

← saying that we want to avoid errors for this subset more...

Cont'd

Want to max the dist. bet. lines. parallel

given by $\frac{2}{\sqrt{\sum a_i^2}}$ (OR) $\text{MIN}(\frac{\text{margin}}{\sum a_i^2})$



SVM optimization: $\text{Min}(\frac{\text{margin}}{\sum a_i^2})$

$$s. [a_0 + \sum_{i=1}^m a_i x_{ij}] y_j \geq -1$$

→ What if cannot hard classify?

Correct side: $y_j (a_0 + \sum a_i x_{ij}) \geq +1$

Wrong side: $y_j (a_0 + \sum a_i x_{ij}) < 1$

Error will happen. How much? No error is 0 (OR)

Distance from classification line

$$\text{So Error} = \text{Max} \left\{ 0, 1 - \underbrace{(a_0 + \sum a_i x_{ij}) y_j}_{\text{amount stayed from line}} \right\}$$

no error

→ Sum of all Errors = $\sum_{j=1}^n \text{Max} \{ 0, 1 - (\dots) y_j \}$

→ Margin dist = $\frac{\text{margin}}{\sqrt{\sum a_i^2}}$

Optimization: $\text{Min}(\text{sum of errors}) + \text{Max}(\text{Margin dist.})$
or, $\text{Min}(\checkmark) + \lambda \sum a_i^2$

This is a tradeoff! Bet. $\sum$ errors and large margins

→ adding multiplier $(m_j \cdot \text{Max} \{ 0, 1 - () y_j \})$

↳ adjust imp. of avoiding error

Scaling Can do this linearly

→ Between $[0, 1]$

for factor x_{ij} ,

$$\text{Scaled } x_{ij} = \frac{x_{ij} - \min x_{ij}}{\max(x_{ij}) - \min(x_{ij})}$$

x_{ij} → factor
i → attribute
j → row

→ Now can extend to arbitrary range $[b, a]$

x_{ij} scaled to $[b, a] =$

$$x_{ij} \text{ scaled to } [0, 1] \times (a - b) + b$$

~~Can~~ Can do this w/normal dist

→ This is standardizing &

$$\text{Scaled } x_{ij} = \frac{x_{ij} - \mu(x_{ij})}{\sigma(x_{ij})}$$

→ Use for data w/ bounded ranges

{ SAT scores
credit scores

{ batting average

???

→ Used in PCA and clustering.

clustering

KNN (Classification!)

~~For~~ You don't demarcate the data points.
You look at the other points that the
data point is in and classify.

$K \rightarrow$ ~~Class~~ The number of "Closest points"
or "nearest neighbors".

Classify

SVM

KNN

How to measure closeness?

simple way is distance. for a new point x_i ,

$$d = \sqrt{\sum_{i=1}^n |x_i - y_i|^2}$$

But some distances can be more important.
So we use weighted distance:

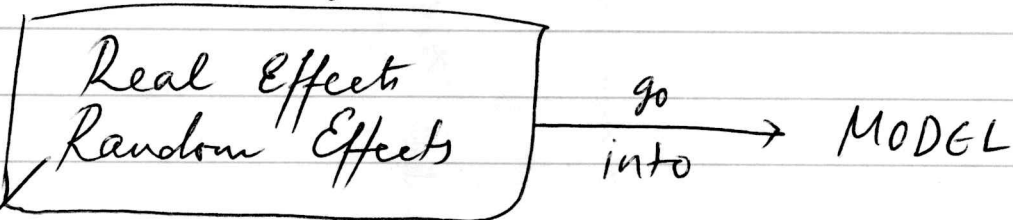
$$d = \sqrt{\sum_{i=1}^n w_i |x_i - y_i|^2}$$

In some cases, $w_i = 0$! Some others
might not be imp. at all!

Validation

Measure effectiveness.

Dataset 1



diff
in
another
dataset!!
Problem!!

When MODEL is applied to diff. dataset,
only the real effects show up.

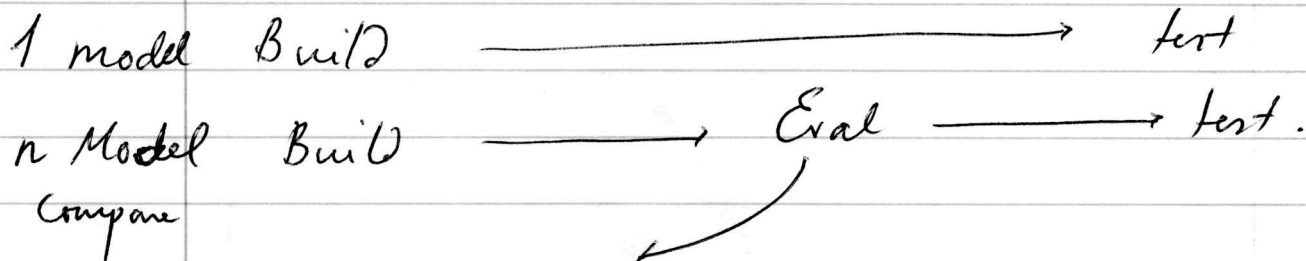
DIFFERENT random effects!

These can make a model look good
when it really isn't.

Training

Validation

Test



high-performing models may get "boost"
from random data.

↓
above avg.
random effects.

→ All about dealing w/ random
patterns in data

How to Split Data

Train - Build
Validate - Compare
Test - Est. perf.

Single Model: 70/30, 90/10
Compare Model: 50/50 split rest
train valid/test.

"It Depends"

How to split? (1) Random! $n = 1,000$

60/20/20

Train = Rand(600, 1000)

valid = Rand(200, 400)

test = 200 (left over)

(2) Rotate each point
between T/V/Test.

problem: Might lead to "buildup" of
data in each set (bias)

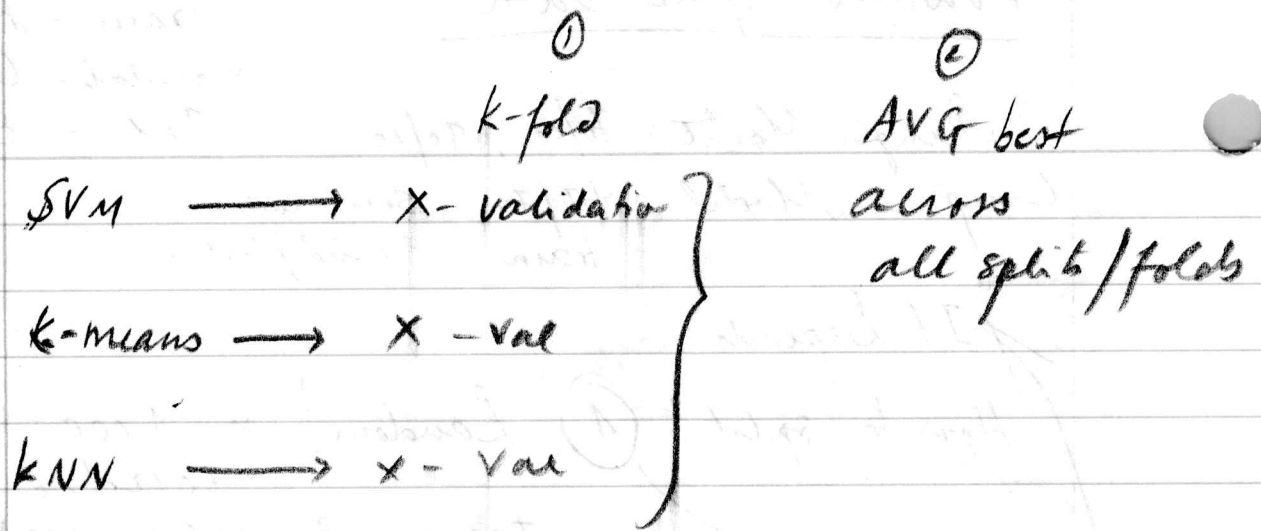
eg: ^{Weekly} Sales. Train gets ~~all~~ MWF only...

problem: May not separate data equally
like with rotation. May cluster
data as well.

eg: Sales over 10 yrs: Train ^{is missing} gets most
of some years.

→ What to do?

also: combine approaches is possible.



③ Pick model w/ best fit (AVG) across all folds

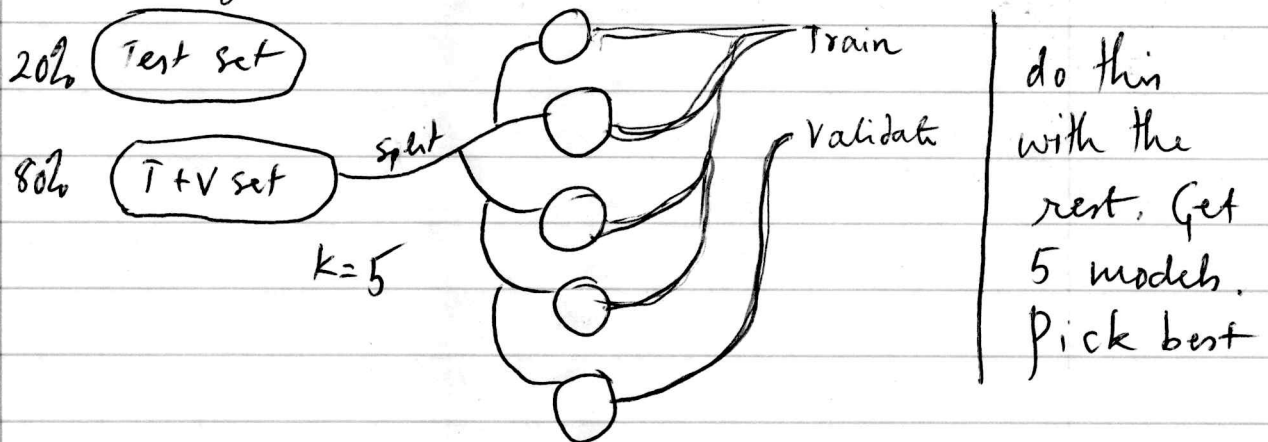
④ RE-TRAIN across all data
DO NOT average coeffs!

K-folds are (Training & Validation!)
NO Test!

Cross Validation

Avoid "build-up" of data points in each bin: Train, Valid, Test.
to avoid bias.

One way is k-fold Cross Validation



→ From this, you will get performance of FIVE models.

What is this? The
AVERAGE across all
"folds" (splits) of
Train. data

→ Now, what to do?

Select model and train again
with all train data.

Do not pick the best avg!

Parameters
Coeff, slope

Hyperparameters -
 C, k

Clustering: group data points.

Eg: used in targeted marketing.

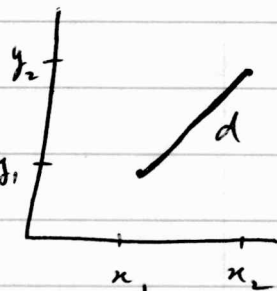
used in medicine $\rightarrow$ response of people to diff. treatments.

used in city planning $\rightarrow$ police station, hospitals.

"content" of something

Norms: (FFS!) (N-dimension?)

$\rightarrow$ Euclidean dist = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
(2-norm)



$\rightarrow$ Rectilinear / Manhattan
(1-norm) dist = $|x_2 - x_1| + |y_2 - y_1|$

$\rightarrow$ p-norm = $\sqrt[p]{\sum_{i=1}^n |x_i - y_i|^p}$
(Minkowski)

Think City grids.

$\rightarrow \infty$ -norm = $\text{MAX} [|x_i - y_i|]$

$\left(\sqrt[p]{\sum_{i=1}^n |x_i - y_i|^p} \right)^{\infty} \approx \sqrt[p]{\text{MAX} |x_i - y_i|^p} = \text{MAX} |x_i - y_i|$

Where does one use this?

Example of warehouse stacking machine

1-norm



$t_x + t_y = t_{\text{total}}$
 $(|t_{x_2} - t_{x_1}| + |t_{y_2} - t_{y_1}|)$

∞ -norm

if it extends arm as it goes along,

$t_{\text{tot}} = \text{MAX} (t_{x_2} - t_{x_1}, t_{x_2} - t_{x_1})$

The big problem you have with this is thinking in many dimensions!

K-Means Clustering

$$\sum_k y_{ik} = 1$$

$$y_{ik_0} + y_{ik_1} + \dots + y_{ik_K} = 1$$

$$0 + 1 + 0 + \dots + 0 = 1$$

What does that mean?

Every point x_{ij} is assigned to exactly one cluster ~~and~~

x_{ij} ~~factor data point~~
factor
a H_r data point

$$y_{ik} = \begin{cases} 1 & \text{if } x_{ij} \text{ in } k \\ 0 & \text{if not} \end{cases}$$

z_{jk} = Coord of Cluster center k

- Goal: ^①Minimize distances to cluster center if and (in English) ^③~~if~~ only if data point is in cluster ^②

$$\text{Min}_{y,z} \sum_i \sum_k y_{ik} \sqrt{\sum_j (x_{ij} - z_{jk})^2}$$

③ ② ①

This is a hard "optimization problem".

∴ k-means $\xrightarrow{\text{"heuristic"}}$

$\xrightarrow{\text{"Expectation Maximization"}}$

(here, maximizing the -ve distance to cluster center)

Algorithm: ① pick

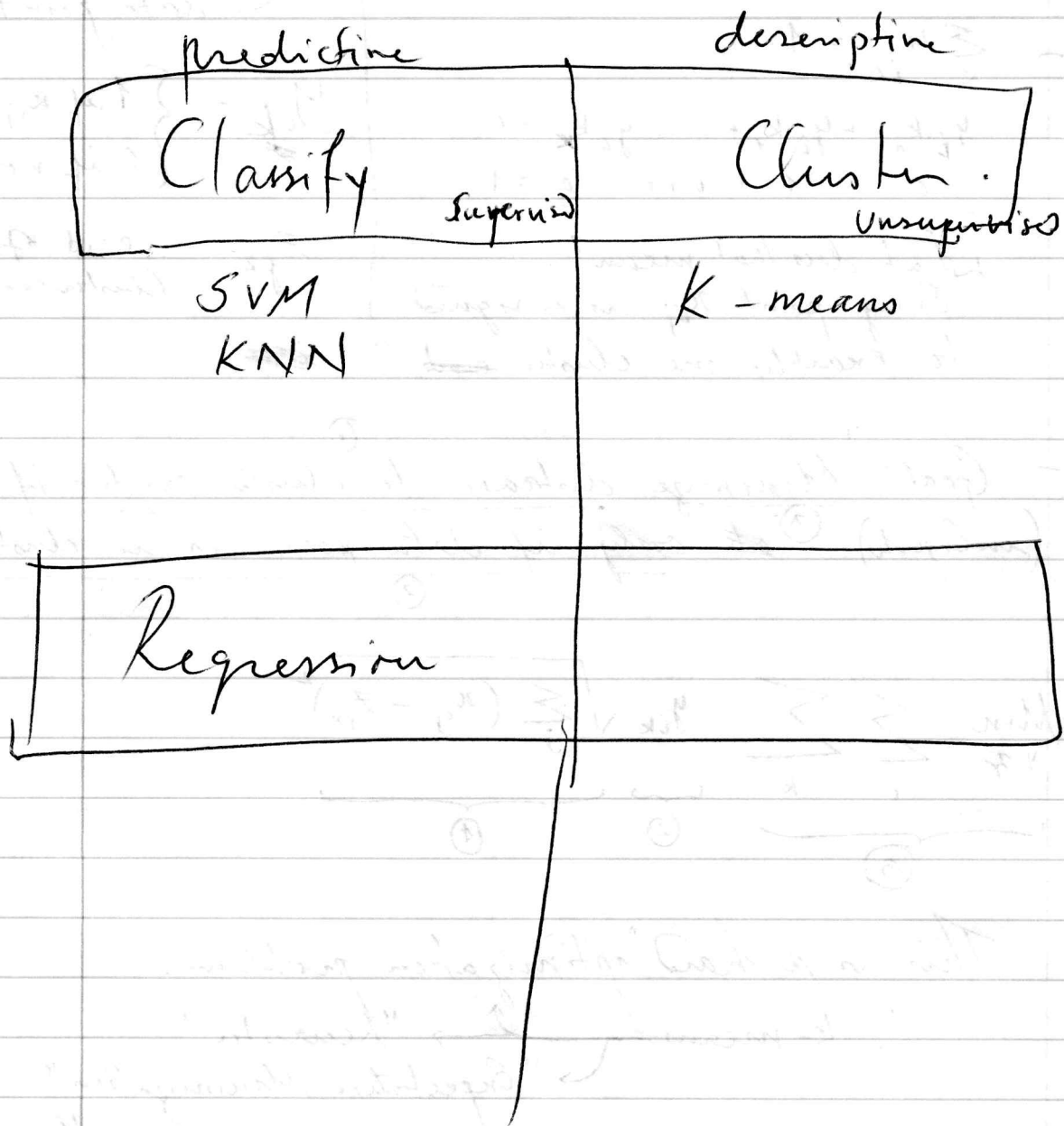
~~assign~~ cluster centers randomly

while no changes:

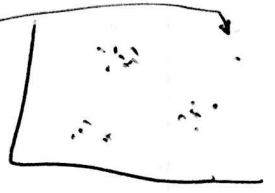
② assign data points to centers (nearest)

compute centroid

assign data to centroid



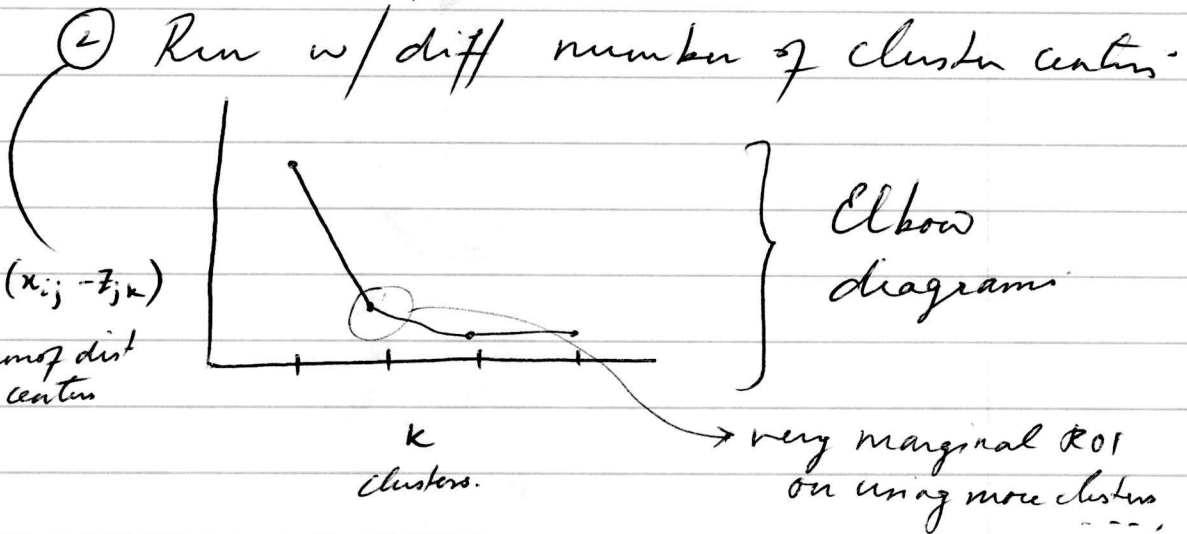
What to do w/outliers?



How to get best clustering?

① Run w/diff. cluster centers

② Run w/diff number of cluster centers



$\Sigma(x_{i,j} - z_{j,k})$
sum of dist
to center

k
clusters.

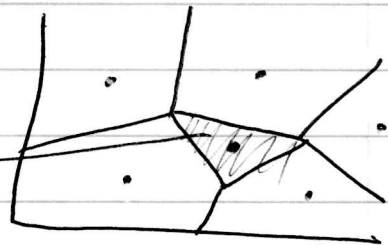
Elbow
diagrams

very marginal ROI
on using more clusters

Use a Voronoi Diagram to show the range
of values for each cluster

"Predictive Clustering"

points closer to • than
other •'s



Supervised

Unsupervised

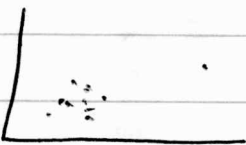
- We know the attributes
and response
- Classification is known

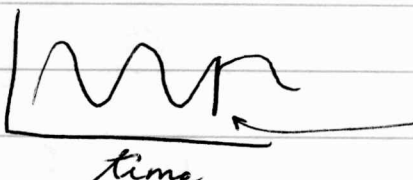
- know nothing about
how data groups
- Class is not known

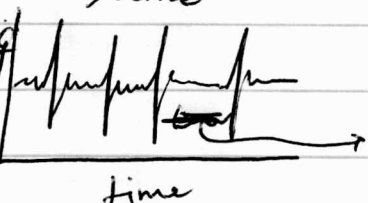
Data Preparation

- Scale data
- outlier detection.

Outlier Detection : There are quite a few types.

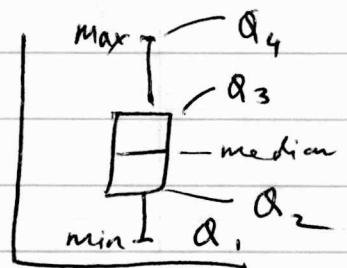
→ point  "far from data"

→ Contextual  "diff. relative to the data"
One thing

→ Collective  many things
missing from
stream of data

Dealing with them : remove bad ones
evidently, verifiably bad.

Whether or not to remove the rest is
"it depends"



Change Detection

Usually done w/ time-series data.
Hypothesis tests might be slow.

- ① Is action needed now?
- ② what happened due to past action?
- ③ Have patterns changed?

CUSUM: Cumulative Sum. ("control chart")
"Has mean gone above some lvl?
or below."

Increases!

$$S_t = \text{Max} \{0, S_{t-1} + (x_t - \mu - c)\}$$

THIS IS NOT ROLLING!!

THAT MEAN μ is for all to adjust sensitivity values!!

Change detected if $S_t \geq T$

↓ threshold
determine empirically

ATBV, for increases,

if c is high → less sensitive
→ less likely to make false alarm

if T is high → takes more time
→ less likely to make false alarm lot

pull down

x_t a little ∴ it
could be a random spike!

DECREASES
 $S_t = \text{max} \{0, S_{t-1} + (\mu - x_t - c)\}$
 $S_t \geq T$

Related to CUSUM? No!

Change detection

Exp smooth: Change to "baseline" due to random fluctuation

Exp. Smoothing: Used for time-series data.

{ Cyclical increase/decrease.

Use to determine baseline response }

because randomness.

Is this something to worry about?

$$S_t = \alpha x_t + (1-\alpha) S_{t-1}$$

S_t $\xrightarrow{\text{Expected Baseline}}$ response at time t

αx_t $\xrightarrow{\text{New Baseline shift it!!}}$

$(1-\alpha) S_{t-1}$ $\xrightarrow{\text{Same old baseline (random luck) keep it!}}$

$0 < \alpha < 1$



{ if α is close to 1, means not much randomness in system
if $\alpha \rightarrow 0$ means a lot of randomness in system.

Start w/ $S_1 = x_1$

Eg pick small α if you think there is a lot of randomness in the system.

No trends or cycles in this simple model "

descriptive use for smoothing and forecasting
Exp. Smoothing → Trends. $T_t = \text{trend at time } t$

$$S_1 = x_1 \quad S_t = \alpha x_t + (1-\alpha)(S_{t-1} + T_{t-1})$$

$$T_1 = 0$$

↑
ADDITIVE

$$T_t = \underbrace{\beta(S_t - S_{t-1})}_{\text{keep new trend}} + \underbrace{(1-\beta)T_{t-1}}_{\text{keep old trend}}$$

→ What about cyclic patterns?
 (also called "Seasonality")

Can do this like Trend (add to formula)
 but let's complicate things.

Inflate or deflate the observed value
 (NOT Baseline + Trend)

$$C_1, C_2, \dots, C_L = 1 \quad S_t = \alpha \frac{x_t}{C_{t-L}} + (1-\alpha)(S_{t-1} + T_{t-1})$$

↑
MULTIPLICATIVE

C_{t-L} → WTF

$$C_t = \gamma \left(\frac{x_t}{S_t} \right) + (1-\gamma) C_{t-L}$$

L is ~~not~~ cycles length of cycles.

Eg: BP measured every hour for a day.
 $L = 24$ (!important)

$C = 1$ no change

$C = 0.9$ 10% lower

$C = 1.1$ 10% higher

Exp Smoothing



2/11/11

$$S_t = \alpha x_t + (1-\alpha)S_{t-1} + (1-\alpha)\alpha$$

$$S_t = \alpha x_t + (1-\alpha)\alpha x_{t-1} + (1-\alpha)^2 \alpha x_{t-2} + \dots$$

- This is "exponential"
- The most recent observation is weighted!

$$S_t = \alpha x_t + (1-\alpha)S_{t-1}$$

if high, pushes it down
if low, pushes it up } "smoothing"

Can use this to forecast things as well

Say forecast is F_{t+1}

BASIC $F_{t+1} = \alpha x_{t+1} + (1-\alpha)S_t$

↳ But we don't know this!!

So best guess is S_t

$\Rightarrow F_{t+1} = S_t$ (but errors can happen due to uncertainty...)

or $F_{t+k} = S_t \quad \forall k=1,2,3,\dots$

TREND

$$F_{t+1} = \alpha x_{t+1} + (1-\alpha)(S_t + T_t)$$

$$F_{t+1} = \alpha S_t + (1-\alpha)(S_t + T_t)$$

$$= \alpha S_t + S_t + T_t - \alpha S_t - \alpha T_t$$

$$F_{t+1} = S_t + T_t$$

$$F_{t+k} = S_t + kT_t \quad \forall k \in \mathbb{N}$$

Derivation will suck ...

SEASONALITY :

$$F_{t+1} = (S_t + T_t) C_{(t+1)-L}$$

$$\alpha, \quad F_{t+k} = (S_t + kT_t) C_{(t+k)-L+(k-1)}$$

$$\forall k \in \mathbb{N}$$

Now how to find α, β, γ ?

This is an optimization problem, lol.

$$\min (F_t - x_t)^2$$

over the training set.

Ex_{mp} Good for short term forecasting

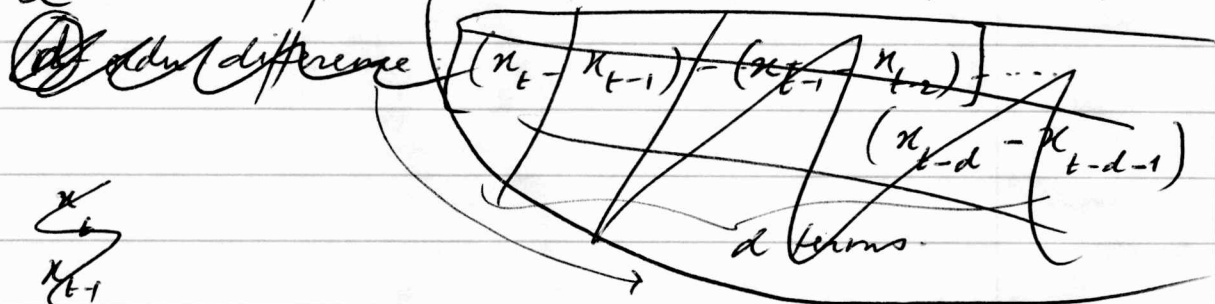
ARIMA time-series analysis (like Exp. Smoothing)

① Differences. ~~But~~ ARIMA deals with data that is not stationary

(This means that μ, σ, σ^2 are constant over time)

Often data is not stationary. BUT, ~~the~~ diffs can be stationary

2-order difference: $(x_t - x_{t-1}) - (x_{t-1} - x_{t-2})$



x_t
 x_{t-1}
 x_{t-2}
 x_{t-3}
3-order diff
 $\text{diff}(\text{diff}(\text{diff}))$

3-order diff is
diff (diff (diff))

$$\left[(x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) \right] - \left[(x_{t-1} - x_{t-2}) - (x_{t-2} - x_{t-3}) \right]$$

~~$(x_t - x_{t-1})$~~

First part is (d-order) diffs

② Auto-Regression :

???

"Regression": predict value based on other factors

"Auto": predict using own, old values.
ONLY WORKS W/TIME SERIES DATA!

$$S_t = \alpha x_t + (1-\alpha)\alpha x_{t-1} + (1-\alpha)^2 \alpha x_{t-2} + \dots \dots \dots \infty$$

This is ∞ -order autoregressive

Say we only use (p) time periods

$$S_t = \alpha x_t + (1-\alpha)\alpha x_{t-1} + \dots \dots \dots (1-\alpha)^{p-1} \alpha x_{t-(p-1)}$$

So far, ARIMA uses (p) time periods (autoreg)
to predict (d) ~~the~~ order diffs.

Auto-regression applied to Differences.

③ Moving Average

ARIMA uses previous errors as predictors.

(No idea) this is (q) order moving Avg
for q time periods.

Put all that together and you get
a complicated equation

But ARIMA is ~~is~~ a good generalization!

$\overset{p}{\text{ARIMA}}(\overset{d}{0}, \overset{q}{0}, 0) \rightarrow$ White noise (no pattern)

$\text{ARIMA}(0, 1, 0) \rightarrow$ Random Walk

$\text{ARIMA}(p, 0, 0) \rightarrow$ Simple Auto-R model
(Switch it on)

$\text{ARIMA}(0, 0, q) \rightarrow$ Simple Moving Average Model

$\text{ARIMA}(0, 1, 1) \rightarrow$ Exp Smoothing!! 😊

This gives ^{forecasts} values. There's another
that gives you ^{forecasts} variance

Time Series

Exp. Smoothing!

ARIMA!

GARCH!

→ When only the previous
values of a response affect
its current value.

GARCH : ~~Est~~ Forecast variance.
Time Series

Why? Generally use to estimate amount of error

How much forecast is high/low than truth

Eg: Forecasting demand for pickup trucks.

Investment portfolios: high or low risk?

variance is proxy
for risk.

Compare to ARIMA:

- { Use variances }
- { Use squared errors }
- { Use raw variances,
NOT diff. of variances.

GARCH(p, q)

REGRESSION

min min ☺

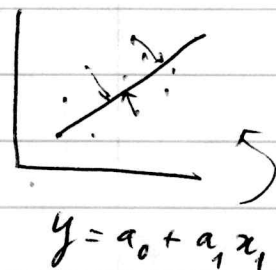
Predictive, Classification
descriptive

How do things work? (Auh....) (descriptive)

What will happen in the future? (predictive)

Predictors $\longrightarrow$ Response

$$a_0 + \sum_{i=1}^n a_i x_i \longrightarrow y$$



Error: Say ^{pred, model} response is $\hat{y}_i$
Say real response is y_i

$$\text{Error} = (y_i - \hat{y}_i)$$

$$\text{S.O.S. error} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\text{Optimization Problem: Min} \left[\sum_{i=1}^n (y_i - \hat{y}_i)^2 \right]$$

for values of $a_i, i \in \mathbb{N}$

Model quality
Why sum of squares for error?

75 80 $\rightarrow$ 8.2%

AIC1 - AIC2 $\rightarrow$ better

e

75 - 60% $\rightarrow$ 12.5%

e

better than ~~it~~

e

Measure Model Quality & Fit (FUUUK!)

(BTW) Simpler models are less likely to overfit data

MLE

AIC

BIC

Max. likelihood Est.

have correct data
know variance.

$$\text{Error} \sim N(0, \sigma^2)$$

$z_i \leftarrow$ ~~param model~~ true

$y_i \leftarrow$ ~~observed~~ model

Each point generated independently
(^{"i.i.d"})

If all things normal, ~~prob~~ prob of seeing

z_i if ~~measured~~ is y_i

$$\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(z_i - y_i)^2}{2\sigma^2}} = PD$$

prob. density function

over all points, $\prod_{i=1}^n PD = \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (z_i - y_i)^2}$

GOAL: ~~Minimize~~ MAXIMIZE

over all params $a_0, a_1, \dots, a_n$.
comes down to min. largest term: $(z_i - y_i)^2$

$$\text{Min } (z_i - y_i)^2 = \text{Min} \left[z_i - \left(a_0 + \sum_{i=1}^n a_i x_i \right) \right]$$

Akaike Info. Criterion

$$AIC = 2k - 2\ln(L^x)$$

This is the penalty term

→ balance likelihood w/
Simplicity.
→ avoid overfitting.

of param
estimates.

ML value

But need as data!

SMALL IS GOOD.

AIC - Corrected: $AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$

COMPARE

AIC1: 75

AIC2: 80

Relative likelihood: $e^{\frac{(AIC1 - AIC2)}{2}} = 8.22$

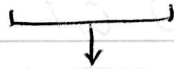
⇒ Model 2 is only 8.22% likely better than Model 1
→ Keep it.

lower AIC is better!

ORDER MATTERS

Bayesian Information Criterion

$$BIC = k \ln(n) - 2 \ln(L^*)$$



L^* : MLE

k : # of param

n : data points

Larger penalty term than AIC!

Encourage models with fewer parameters than AIC

"only use when more data points than param"

> ???

Rules of Thumb!

$$|BIC_1 - BIC_2| > 10$$

Smaller BIC model "very likely" better

$$6 < \dots < 10$$

"likely"

$$2 < \dots < 6$$

"somewhat likely"

$$0 < \dots < 2$$

"slightly likely"

Regression: Data Transformation

Take something as plain -

$$y = a_0 + a_1 x_1 + \dots + a_m x_m$$

Transformation

Point is that you can change the response
change the attributes/predictors

$$\text{eg: } \log(y) = a_0 + a_1 x_1^2 + a_2 \sin(x_1^2) + \dots$$

Variable

Interaction

Or you can club two predictors together as well

$$\text{eg: } \log(y) = a_0 + a_1 x_1 + a_2 x_2 + a_3 (x_1 x_2^2)$$

Considered a new col. of data

Regression: Reading Output ♥♥♥!

$$y = a_0 + a_1x + \dots + a_mx_m$$

p-values: Prob. that coefficient is zero.

Hypothesis testing

usually, if $p > 0.05$, remove
higher means more predictors
more error

lower means fewer predictors
more error

Commonsense
Usual Caveats: ① May mean nothing even
when they are very low

i.e., not related
to the response

② They're prob. may not
be right even when low.

$n = 1000$, $p = 0.01 \Rightarrow$ ~~10 are~~
~~irrelevant~~

Each has a 1% chance of
not being sig.!

10 are irrelevant (Expect)

Confidence Interval: Where the coefficient prob.
lies (eg. 95% conf. Int.)
& how close it is to zero

t-statistic: $\frac{\text{p-value}}{\text{standard error}}$

??

Just see Coefficient : if you multiply it by predictor value, and it doesn't make much diff.... ✗

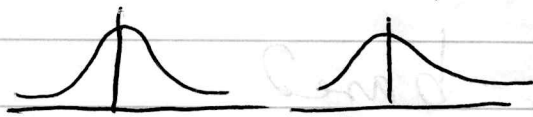


R² Value : As a whole, how much variability does your model account for? The rest is random (or something else)

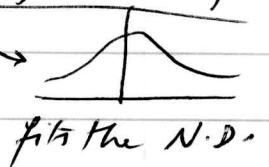
eg: $R^2 = 0.54 \Rightarrow$ Model acc. for 54% of variability.
w/ 46% being random or other shit.

Advanced Data Prep

Not all data fit the Normality Assumption



more variance.

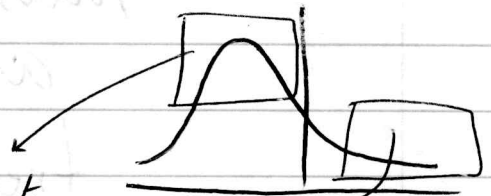


fits the N.D.

Heteroscedasticity: Unequal variance.

Box-Cox: Transform a vector of responses to come close to ND

$$t(y) = \frac{(y^\lambda - 1)}{\lambda}$$



enlarge to increase variability

~~enlarge~~ shrink to reduce variability

How to check this?

Q-Q plot.

Variance!

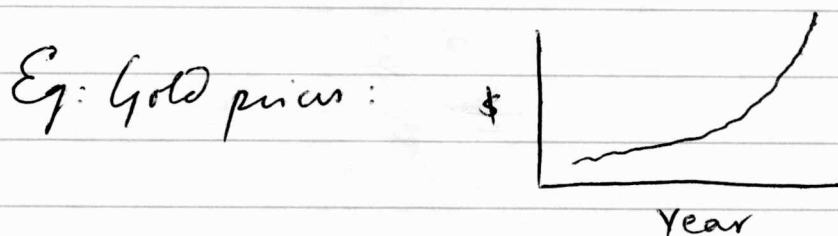
Factor-based
Regression, SVM

~~are not sensitive to~~
~~must be detrended~~
are not sensitive to
trends, seasonality ...

(hence detrending)

Detrending Data (Advanced Data Prep)

"Trend": data increases or decreases over time.



however, what about inflation??
how to remove this?

Go factor by factor in your model and fit one-dim. regression to it.

$$\underset{\text{trend}}{\$} = -43,200 + (23.2 \times \text{year})$$

Now remove this trend from the real price.

$$\underset{\text{detrended}}{\cancel{\$}} = \$ - \underset{\text{trend}}{\$}$$

This actually, at least in the gold example, comes pretty close if you cheated and looked up inflation rates and adjusted the original graph...

Use for factor-based models like Regression and SVM that don't know about "trends" in data...

PCA

(Advanced data prep)

"Which factors do I use?"

lots of factors. Which ~~do~~ are important?

- Have lots of predictors
- Data for predictors might be irrelevant, changed.
- Predictors might be highly correlated.

PCA: Removes correlation

Ranks coords by importance
Reduces randomness.

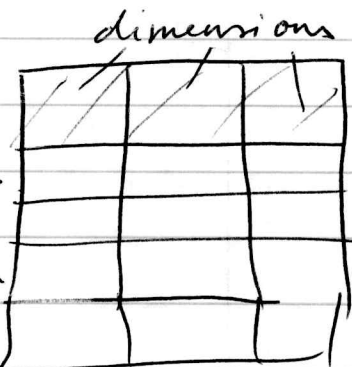
First few coords will have high signal-noise.

$$X = \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}$$

- Scale data
- Normalize Mean
- Find eigenvalues and Eigen vectors.

$$\frac{1}{n} \sum_{i=1}^n x_{ij} = 0 = \mu_j$$

factor 1
factor 2
factor 3



"Which way is the scaling data moving the most?"

• Get new factors

of PCs = whichever is smaller.

• Each new factor is linear
Combs of original factors

Can use kernels, not
just linear transformation!

$$\begin{bmatrix} X \\ \text{data} \end{bmatrix} \begin{bmatrix} \sqrt{\lambda} \\ \text{eigen vector} \\ \text{matrix} \\ \text{sorted by eigenvalue} \end{bmatrix}$$

$$XV_1 = \text{first PC} : PC1$$

$$XV_2 = \text{second PC} : PC2$$

$$k^{\text{th}} \text{ factor for } i^{\text{th}} \text{ data point} \\ = t_{ik} = \sum_{j=1}^n x_{ij} \cdot v_{jk}$$

→ How to turn back to original factors?

Have to plug in t_{ik} and do the math "

$$\text{But } a_j = \sum_{k=1}^L b_k \cdot v_{jk} \quad \begin{cases} \rightarrow L \text{ factors total} \\ \rightarrow b_k \text{ is coeff in 'new' model.} \end{cases}$$

THEN, use regression to find NEW coeffs $b_0, b_1, \dots, b_L$

$$y_i = b_0 + \sum_{k=1}^L b_k \cdot t_{ik}$$

$b_0, b_1, \dots, b_L$ ← total

(Youtube: StatQuest PCA
Step-by-step)

CAR Tree (Regression)

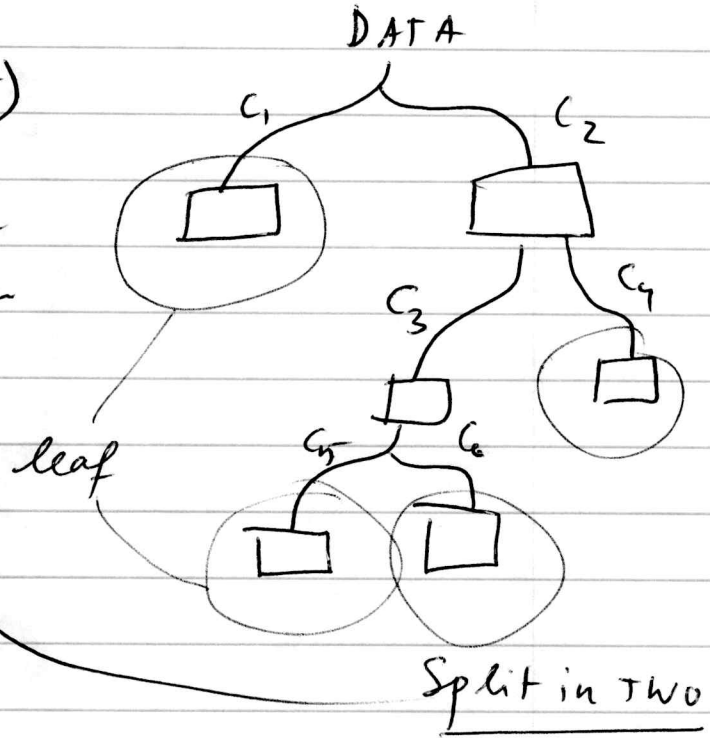
Classification ~~diff~~ diff

Decision Maker's
~~tree~~
Decision trees

Classification
and Regression
Tree.

C_n = Criterion
("Age > 25")

Each leaf has a
different regression
model.



How to branch?

{ Which factor?
how to split?

for each (leaf):
for each (factor):

① Take half the data

② Run regression model on it.

first leaf!

For ~~each~~ each leaf

For each factor:

{ Split on each factor

{ Get variance of both leaves

(TOTAL

compare to threshold

{ Split for biggest decrease

Choose factor
with lowest
total variance!!

③ Use other half to prune.

???

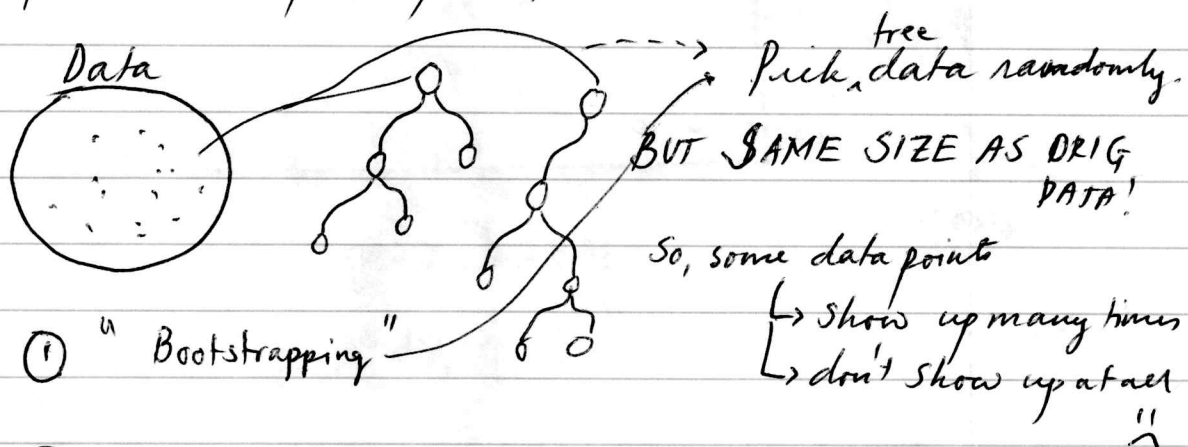
for each branch point:

get estimation error with and w/o branch
prune if this gets larger

ALSO: Overfitting is a prob. Rule of Thumb: Each leaf has at
least 5% of data points

CART: Random Forest

Generate a forest full of trees! Randomly?



② Branching: Pick a random set of factors
(common: $1 + \log(m)$)

Pick a decent factor to branch on.

③ Pruning: LOL WUT! No pruning.

Now you're ready. Run!

If REGRESSION TREE: Average the response

If CLASSIFICATION TREE: Most common response
(decision tree?) (the mode)

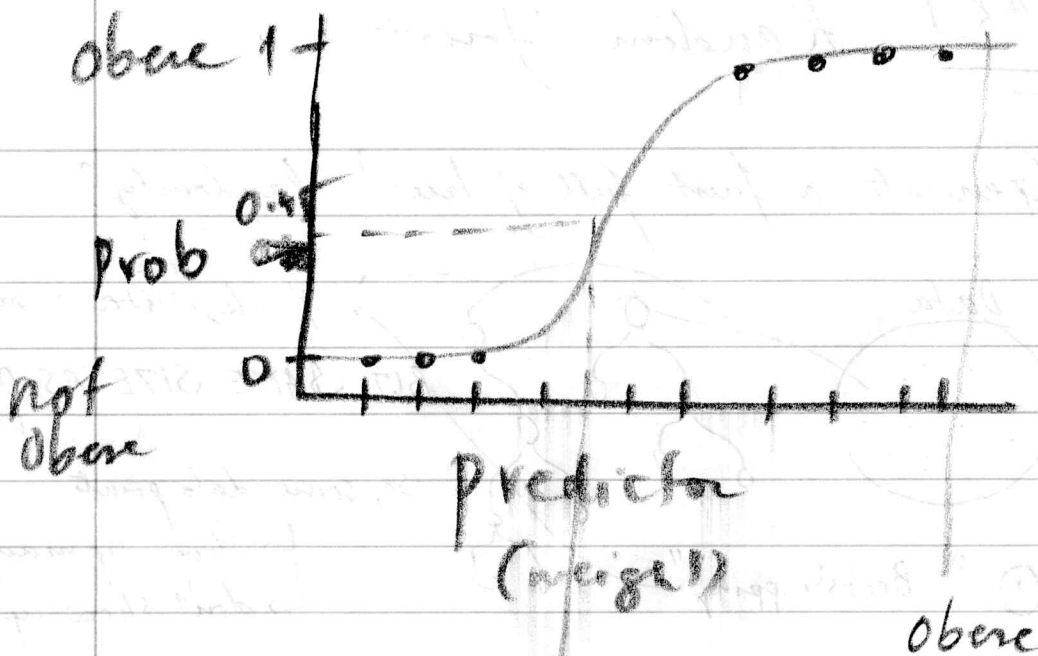
Good Shit

- Better overall estimates
- overfitting problem is "Eh" because of averaging

Bad Shit

- Hard to explain
- Hard to interpret
- There's no specific model you can point to....

Easy AF so called "default model" lol.



"Summarizes Confusion Matrix!"

45% chance of obesity

↑

ROC

pick threshold that makes this happen

% correct class > % incorrect class

* Stat print on YouTube

Sen
(TP rate)

any point on this line ⇒

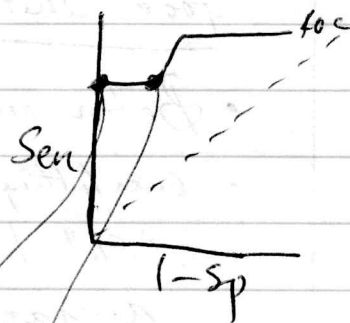
% of correct classification

SAME as % of incorrect classification

1-Sp
(FP rate)

Don't pick this....

→ Pick this!



Better! Small FP rate (zero!)
Bad, non-zero FP rate in comparison!

Regression : LOGISTIC !

Good for two things : Response is Probability val [0, 1]

Response is Binary.

Just simple change to Linear Regression eq.

↓
(Response is Continuous)

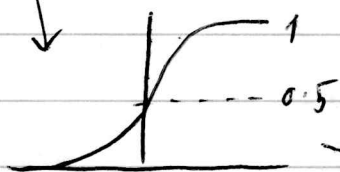
Linear: $y = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_m x_m$

Logistic: $\log \frac{p}{1-p} = a_0 + a_1 x_1 + \dots + a_m x_m$

$$\Rightarrow p = \frac{1}{1 + e^{-(a_0 + a_1 x_1 + \dots)}} y$$

That gives values between 0 and 1 (Probability)

gives 0 if $y = -\infty$
1 if $y = +\infty$] (Binary)



→ The curve is affected by coefficients

- You can do everything with Logit R as with Lin R
- BUT! Take longer to calculate
"No closed-form solution" ???

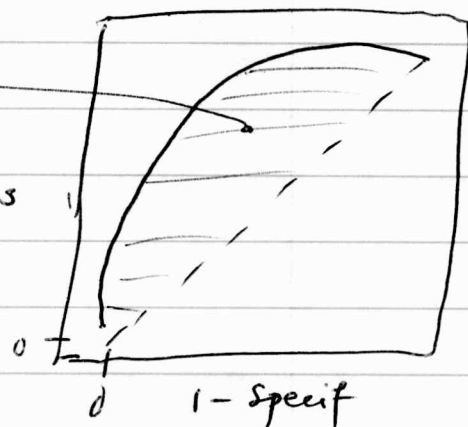
Linear has R^2 : frac. of variance explained by model

Logistic has pseudo- R^2 : not really explaining
but metr.

Regression: Logistic

Use for classification. if $p \geq 0.7$ ^{loan, eg.} do, else don't.
To do this shit, make ROC curve.

→ Estimate prob. that
model will give high p
to one from YES group
than one from NO group.



→ This is Area Under Curve (AUC)
or "Concordance index"

→ if $AUC = 0.5$, This is a fucking GUESS

→ This is quick 'n dirty.

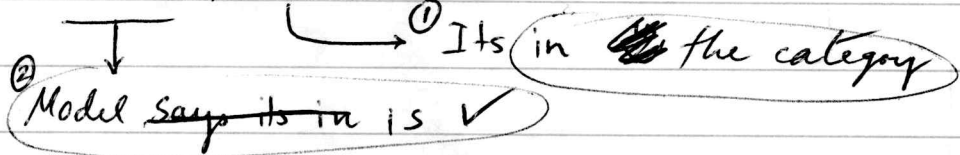
No idea of cost of False -ve, False +ve.

CONFUSION MATRICES

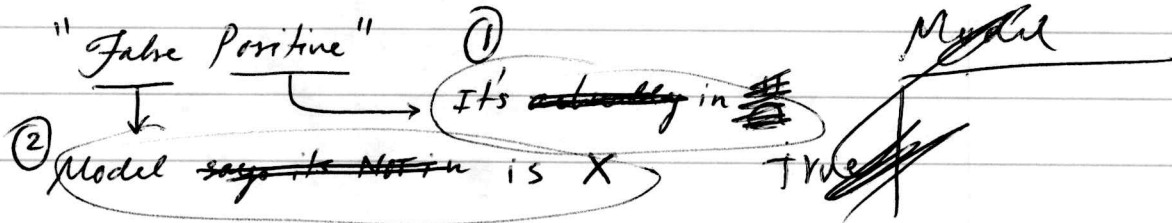
Classification
Evaluate models like

- SVM
- KNN
- Logistic Regression.

"True Positive"



"False Positive"



So, total would be sum of all positives!

Email Example

"In Category" = "is Real"

"True",
"Real"

	Model	
	True	False
POS	TP	FP
NEG	TN	FN

WTF

$$\Sigma \text{ ALL Email} = TP + FP$$

$$\Sigma \text{ Real Email} = TP + FN$$

	MODEL	
	Real	Spam
Real	TP	FN
Spam	FP	TN

True/"Real"

misidentified

all real

Sensitivity : $\frac{TP}{TP + FN}$ → Category members correctly identified

Specificity : $\frac{TN}{TN + FP}$ → Non-Category members correctly identified.

	Model	
	Yes	No
True Yes	✓	X
True No	X	✓

"True negative" : Its not in Category
→ Model is correct.

Validation,

Over fit $\Rightarrow$ worse perform on test

FUCKING SHIT

use this for cost of misclassifying.

WTF is a key,

Poisson Regression : Use if data follows a Poisson Dist.

Eg: Count arrivals at airport line.

$$f(z) = \frac{\lambda^z e^{-\lambda}}{z!}$$

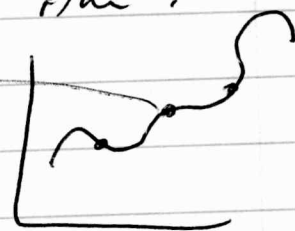
not a single param.
it can be a function of other attribute
(eg: time)

you estimate $\lambda(x)$

Regression Splines

polynomials that connect to each other.

They connect at "knots" must be smooth.



Order-k regression spline: polys are all order k.

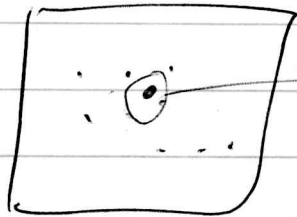
Bayesian Regression Use when you don't have much data to start with.

Start with some idea of $\left\{ \begin{array}{l} \text{coefficients} \\ \text{Random error} \end{array} \right.$

and go from there I guess.

(OR) Use a "broad ~~flat~~ ^{prior} distribution" as "seed data".

KNN Regression : Easy AF
No pred function.



New point?

Take avg response of the
k nearest data points
and be done with it.

EXTRA CREDIT

- Can weight dimension of distance to neighbor
- Remove dimensions

Week 8

Optimizing slides Variable Selection

Why?

①

Avoid Overfitting

- if # of factors the same as # of data points.

- Random effects are given priority! More of a head.

Correlation

(V3)

Causation.

$p > 0.05$

⇒ 52% chance it is not meaningful

420, 64% chance of including a meaningless factor

How?

$$P(\text{No Sig. results}) = (1 - 0.05^{20})$$

$P(\text{at least one sig. result due to chance})$

$$\begin{aligned} &= 1 - P \\ &= 1 - (1 - 0.05^{20}) \\ &= 0.64 \end{aligned}$$

what is this?

②

Prioritize Simpler models

- Better than complex ones.

- Less chance of insign. factors affecting outcome.

- Easier to explain ← important!

③

Illegal to use certain factors

- age, sex, etc for credit decisions.

- Take them out?

NO: Have to remove other factors that correlate highly with

How? Variable selection in context of Regression (others possible)

GREEDY

- ① Look at one thing at a time
Step-by-step
- ② Future options not considered.

Things to use to filter factors

- p-value
- R^2
- Akaike (AIC)
- BIC

Forward selection

- ① add factor start with no factors.
- ② Add gradually.
 - Is it "good enough" ($p \leq 0.05$)
 - Do we "have enough"?

③ fit model

④ Remove factors w/ $p > 0.05$

Backward selection

① Opposite. Start w/ all factors.

② Remove gradually

Is it "bad enough" ($p > 0.05$)

Do we have enough?

③ fit ④ Remove factors.

Stepwise Regression

- Combine both!

Here - Good enough? $\rightarrow$ fit $\rightarrow$ Bad Enough? $p > 0.05$

- Remove $p > 0.05$
- fit model!

Have enough! $\leftarrow$

GLOBAL

(optimization)

xxx

Need to scale the data!

- We are giving a "budget" to the model to spend on co-efficients.
- Don't want one to eat all the budget!

LASSO Regression

$$\min \sum_{i=1}^n (y_i - (\sum_{j=1}^m a_j x_{ji}))^2$$

$$\text{s.t. } \sum_{i=1}^m |a_i| \leq \tau \rightarrow \text{budget!}$$

How to pick?

- ① How many variables do you want?
 - ② Quality of model?
- Pick diff. values of τ

Elastic Net

Same, but constrain
L_{ABSVAL}, AND
L_{sum of squares}.

$$\text{s.t. } \lambda \sum_{i=1}^m |a_i| + (1-\lambda) \sum_{i=1}^m a_i^2 \leq \tau$$

Same way to select τ, λ

Ridge Regression

Just take out the abs val.

$$\text{s.t. } \sum_{i=1}^m a_i^2 \leq \tau$$

Does not do variable selection!

COMPARISON

- Greedy are okay for initial analysis
- ~~Don't~~ Fit Random Effects more.

Elastic Net = LASSO + Ridge

$$\sum |a_i| \leq \epsilon$$

Good for variable selection

Takes out coeffs (makes them zero) to simplify.

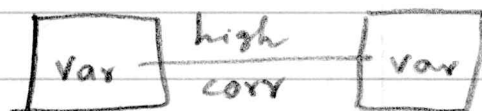
$$\sum a_i^2 \leq \epsilon$$

Good for prediction

Reduces variance in estimates. Shrink coeffs to zero but ~~&~~ includes a lot more coeffs.

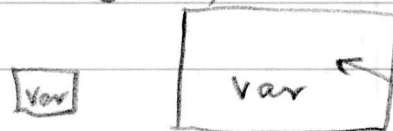
- Prediction Error = f(Bias, Variance)
- Trade off bias for variance for better prediction model.

What sucks? Lasso:



Will just pick one !!

Ridge:



Will push them closer to each other (∵ quadratic)



~~good~~ predictive power is lost !!

DESIGN of EXPERIMENTS

Why?

- Full set of data is not always there.
- Which ad/design works best?
 - Which related product on Amazon?
 - Cannot survey everyone: Representative sample?
 - Compare treatments

Comparison and Control:

Compare two factors (or more) by controlling for others.

Blocking: A factor that creates variation.

Eg. \$ Cars: Do Red or Blue sell more?

- Control for age, type.

- Family vs Sports in a blocking factor.

Lower σ^2 if you compared family red and sports red than if you lumped them together.

Binomial (Two alternatives) A/B Testing Eg. Danner Ads.

- Show option (ad) Randomly
- Collect data
- Run hypothesis test to find option
AFTER EVERY DATAPOINT COLLECTION!

Only use when:

- ① Can collect data quickly
- ② Data ^{sample} representative of population
- ③ Can arrive at decision with a small amt. of data

Eg: Expect 2500 clicks but
Wast 2000 collecting data !!

Several alternatives?
Several factors?

Factorial Design: Why one ad and not another?

- Font x 2
 - Color x 2
 - Background
- } 8 combos to test
- FULL FACTORIAL $\Rightarrow$ test all

What if there are many combos?

Use FRACTIONAL FACTORIAL DESIGN

Take a subset of combos.

Use Balanced Design : ^{TEST} Each choice same #
(see slides) of times.
(software does this for you...) Each pair of choices
Same # of times.

If you have (or think you have)
independent factors, use FRACTIONAL
with regression.

eg: Response ← Model with
(Clicks)
 { Bg color (White, Red)
 { font size (10, 12, 14)
 { font (Arial, etc)

May include
interaction terms → white font on
white bg " "

Exploration vs Exploitation

"How quickly are you able to
determine value with your tests?"
(models, etc)

Gather info Use it

Data is a fixed, finite resource
This is a tradeoff.

Testing Alternatives?

{ A/B

{ Higher-order A/B
Multi-Armed Bandit

Testing Combos?

{ Factorial { Full
Fractional

Multi-Armed Bandit: Deal with Exploration/Exploitation problem.

- Name comes from slot machines.

One-armed bandit.

the lever

prob of winning
+ the wins

} Calibrated for "negative exp. val"
(you will lose money on avg.)

So have a ton of slot-machines and
① begin by playing them all. (MAB!)

GOAL: Find machine w/highest Exp. payout.

- ② Then test to find ones w/good
- ③ Update prob.
- ④ Test until best is clear.
- ↓
old Ladies in Vegas....

Exploration

We are testing many alternatives

Exploitation

Making it more & more likely to pick best

We are doing both!
(Simultaneously!)

PROBABILITY-BASED MODELS

Bernoulli: Will x happen or not?

Single Event (Bernoulli trial)

Prob Mass Func.

$$P(X=1) = p$$

$$P(X=0) = 1-p$$

More useful when you ^{chain} ~~clump~~ them: some series of events.

→ Bernoulli Random Variable w/prob p .

Binomial: x successes out of n ^{Bernoulli} iid trials.

Prob. Mass Func

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

for large n , this is a normal dist ☺

Geometric: How many Bernoulli trials until...?

PMF $P(X=x) = (1-p)^x p$ → no. of failures until success $(1-p)$

or $p^x (1-p)$ → success p

Careful. p could be a "success" or "failure" } Context!

⇒ Awesome ∴ use to check iid!
 ↳ if fits Geometric, it is!

Bernoulli

Coin flip

Binomial

of ~~flips~~ ^{heads} in n flips

Geo

of trials/failures before heads

defined as "success"

iid!!

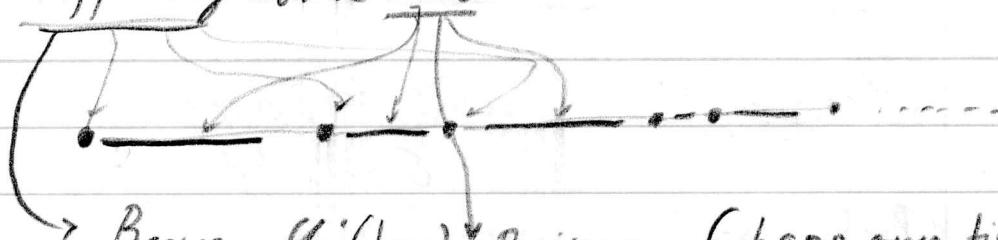
is important!



data cannot clump!

Eg temp data over a year.

happening over time



Bernoulli (happ) Poisson (happ over time)

Binomial (happ) Exponential (time)

Geometric (happ) & Weibull (time)

ii) Poisson : x things happening with average happening rate λ

PMF
 $f_x(x) = \frac{\lambda^x e^{-\lambda}}{x!}$

Model random arrivals of things over a time period

Exponential : If, then happening rate λ is/follows exponential dist

PMF
 $f_x(x) = \lambda e^{-\lambda x}$

AND

Special case of

If inter-arrival/happening time is/follows exponential, arrivals follow Poisson

Weibull : Amount of time until x^a happening (dike failure)

$\left(\frac{k}{\lambda}\right) \left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x}{\lambda}\right)^k\right)$

Shape param

Scale param

AND time between failures

Geometric : # of trials between... (failure)

Weibull : time between... (failure)

(light bulb example.
 # of flips until fail? Geo
 amt. of time until bulb out? Wei)

$k < 1$

→ "Worst things fail first"
 → failure rate decreases w/time

$k > 1$

OPPOSITE

Fewer fail at first
 more over time } eg. fires, bulbs.

$k = 1$

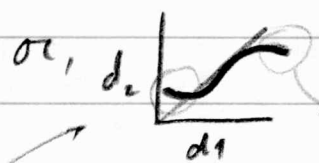
↳ failure rate constant $\Rightarrow$ Exponential!

Q-Q Plot

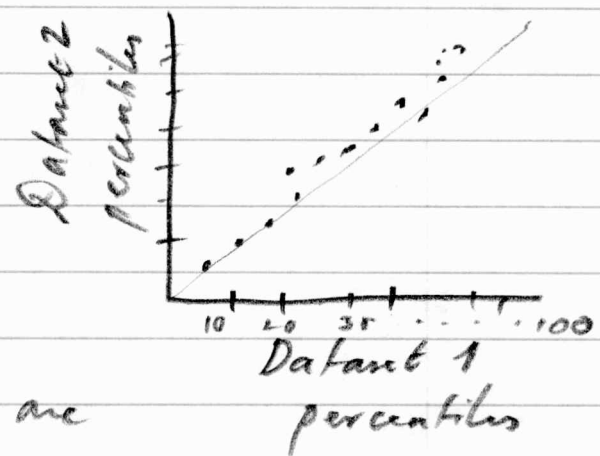
Help see if two datasets have the same dist. and how they are similar

nth-percentile
"larger than n% of data, and
smaller than $\frac{(100-n)}{100}$ of data"

② Does this by comparing percentiles! (or quantiles....)



tail ends are not the same.

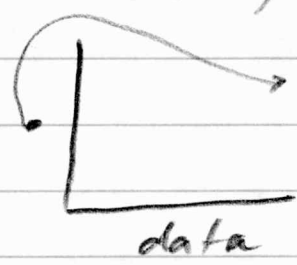


Stat Test can tell us they are not a good match.

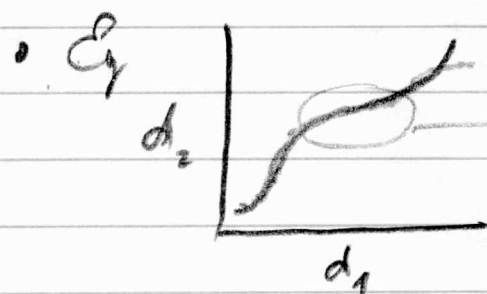
But does not give that visual info

→ If studying Extreme values, test is leading us down the wrong path.

① See how they fit prob. dist.

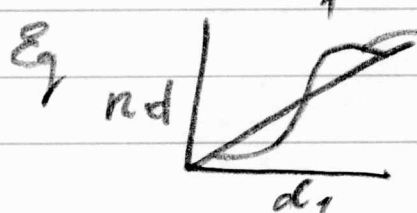


Theoretical value from dist at this quantile / percentile



Memorize? Fuck that

data in d_2 is squashed towards the middle



More points at tails than n.d.

Queuing Memoryless property:

(Call center Example)

- Doesn't matter what happened in the past

[It fits Exponential] - Related to iid?

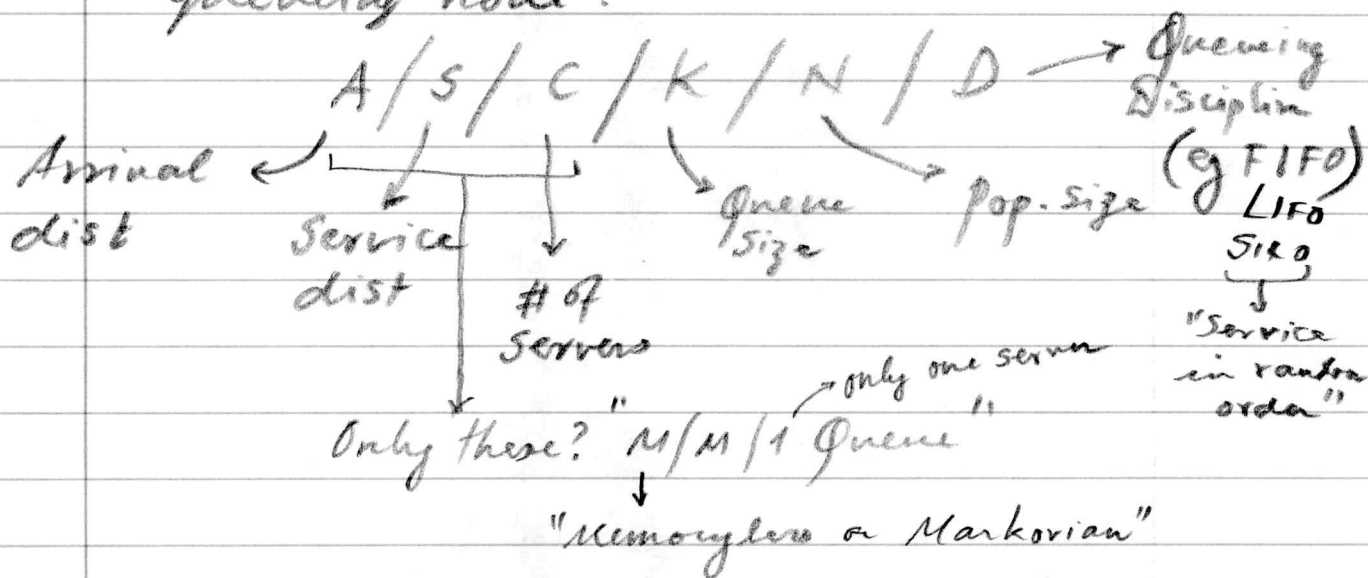
interarrival time fits
 ⇒ Poisson dist for arrival prob.

- Exp. and Poisson
- State depends only on current state.

Not Memoryless

- Eg. bulb, tire failure
- Maybe use Weibull ($k > 1$, $k < 1$)

This is all Queuing Theory. Kendall Notation is standard way to classify a queuing node.



A Queuing Model requires some semblance of a functioning queue, yo. !!

Simulation : - Not all things, sadly, will neatly follow a prob. dist. !!
- Build model and watch behaviour using computing

Deterministic
no randomness

Stochastic Simulation

{ Continuous-time sims (PDES are used)
Discrete-Event sims.

Discrete-Event Stochastic Sims

① Build a model

- Entities (what is moving through sim? bags? plans? people?)
- Modules (queue, storage)
- Actions
- Resources (guards, ~~workers~~)
- Decision points

② Replication : Run many times to get dist of outcomes.

③ Validate : Use real data for this.
look at AVG and Variance!
Can have same avg and diff variance!

④ Use! Simulate w/ diff options (eg: more baggage checks)
Then compare.

Be Careful! - Diff. random numbers can lead to better results.
- Use the same RNs for sims.

GIGO, like anything else.

You are essentially
looking for an
Eigenvector with an
Eigenvalue of 1

Two curved arrows are drawn on the right side of the page. The top arrow starts near the top right edge and points down and to the left towards the word 'essentially'. The bottom arrow starts further down on the right edge and points up and to the left towards the word 'Eigenvector'.

Markov Chain Models

Based on
State of system.

p_{ij} = transition prob of going from State $i \rightarrow j$
 $P = \{p_{ij}\}$ = transition matrix

Start with initial
State vector, π

$$\pi = [p_1 \dots p_n]$$

$$\pi \cdot P = [p'_1 \dots p'_n]$$

System is
in this state
after one
transition

P	S_1	S_2	...	S_n
S_1	p_{11}	p_{12}	...	p_{1n}
S_2	p_{21}			
$\vdots$	$\vdots$			
S_n	p_{n1}			p_{nn}

gives prob. after one state transition.

$(\pi \cdot P)P \rightarrow$ after two state transition.

$\pi \cdot 1^\infty \rightarrow$ yeah... no

Determine "Steady State" $\rightarrow$ Transitions
don't change after a while!

$$\pi^* P = \pi^*$$

So, solve for π^* s.t. $\pi^* P = \pi^*$

$$\text{and, } \sum_i \pi_i^* = 1$$

"Use when
Memoryless
is imp. in
Short run
but not in
long run."
??? WTF?

$\rightarrow$ Memoryless only!

$\rightarrow$ No cycles!

$\rightarrow$ Every state reachable from the others!

Eg: PageRank: Best webpage for topic?

$p_{ij} \propto \begin{cases} P & \text{if } i \text{ links to } j \\ 0 & \text{if not} \end{cases}$ | Find steady state
and rank.

Hint: System is memoryless.

WTF : x_{ij} $\rightarrow$ j th factor in
datapoint i

//
what does this look like??

MISSING DATA and IMPUTATION

- This is a fact of life. Stuff will be missing or be wrong. That's that.
- Some are more likely to be missing than others

What to do when shits' missing? BIAS in model!

- ① Throw away! + Easy
YOLO Not introducing error
- Might lose too much data
Missing data might be biased (eg. rich people not reporting... failed heart transplants....)
This affects the model!!

② Categ. Vars to show missing!

ID	height	kids	ID	height	Missing?	kids
1	20	1	1	20	NO	1
2	40	4	2	40	NO	4
3	?	3	3	0	YES	3
4	30	1	4	30	NO	1
5	15	2	5	15	NO	2

→ indicates missing

But maybe missing is biased? Might be more likely to have kids! ADD interaction terms / vars!

ID	height	missing	kids	kids when height not missing
1	20	N	1	1
2	40	N	4	4
3	0	Y	3	0
4	30	N	1	1
5	15	N	2	2

THESE TWO DO NOT ESTIMATE!

Ramani
Ramani

???

- You are doubling the other vars (kids)
 - If we include interaction terms, we are creating two models, one when there's missing data in "kids" and one when there is not
- with Cat. Var

Imputation: "Estimate missing values".

SIMPLE Missing Data { Numeric: Mean, Median
Categorical: Mode.

⊕ → Easy
→ Won't be "too wrong"

⊖ → Bias is a big problem

PREDICTIVE
(eg. Regression)

⊖ → Tries to address Bias

Complex "

Use data twice: impute, then fit " (OVERFIT)

Does not capture variability.

eg. rich people unlikely to fill survey about income.
Mean will underestimate

What to do? Add some perturbation to each imputed val
Random amount up/down based on, say, the Normal Dist!

NO DISTURB

Average ☺
Variability ☹

DISTURB

Average ☹
Variability ☺

(based on actual variability)

Don't do ~~more~~ ^{a lot} of imputation, tho.

Rule of Thumb $\rightarrow$ not more than 5%
 $\rightarrow$ Use other approaches after this

Errors: So now, ~~errors~~ so far

$$E = (\text{Impute Error}) + (\text{Perturb Error}) + (\text{Model Error})$$

Ya okay but data itself has errors!

Don't worry too much, yo.

Last: M I C E

(
(
(
Equations
Chained
Imputation
Multivariate
)
)
)

Can impute
multiple factor values
for a datapoint
together.

ADVANCED MODELS

What if we don't have good models for the distribution? Hypothesis tests assume this... but what if there isn't?

Unknown Distribution? Non-Parametric Tests

McNemar's Test: Use Binomial to compare pairs of responses in matched results

Treatment Example: A works 61/100
B works 68/100

<u>Scenario 1</u>	<u>Scenario 2</u>
32 : !A or B	12 :
61 : A and B	41 :
7 : B and !A	27 :
0 : A and !B	20 :

B ($p=0.02$) better

B ($p=0.38$) not better

"2% chance
that
pure luck"

Throw out all ^{Cases} ~~examples~~ where results same

What about the ones where responses are not same ("discordant")? Is something going on?
→ then reject NULL HYP

Use Binomial to tell if just by luck.

on Youtube, use χ^2 Test, $\chi^2_{1,0.001} < P < \chi^2_{1,0.01}$
of freedom

Non Param

Not much data

Need median

Dont know distribution

Wilcoxon Signed Rank: Assume dist of response is continuous and symmetric.

"Is median different from some val m ?"

Response y_i . Then • Rank $|y_1 - m|, \dots, |y_n - m|$ from small to large

SINGLE

• $W = \sum \text{rank}(y_i - m)$

$y_i > m$

Sum all ranks where $y_i > m$.

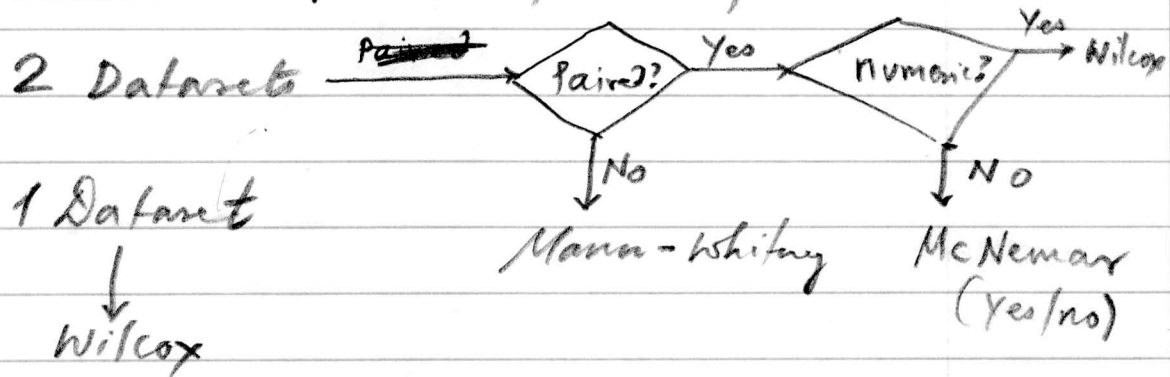
• p-value test. to say median is diff.

PAIRED

Instead of $|y_i - m|$, use $|y_i - z_i|$

Measure: how similar are they? No need for distribution!
 diff. responses

Mann-Whitney: Use for unpaired datasets.



Bayesian Models

Cond. prob. $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

of false +ves lead to weird, counterintuitive results

Empirical Bayesian Modeling:

- Overall dist is known or estimated
- Very little data.

HOLY SHIT WTF?

- Take single observation, combine ^{relate to} with larger observation, make predictions.
- Combine with the expert opinion (say) to get $P(A) \leftarrow$ "prior dist"
to get $P(A|B) \leftarrow$ "posterior dist"
(revised dist)

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

$$\left[P(A) P(B|A) + P(A^c) P(B|A^c) \right]$$

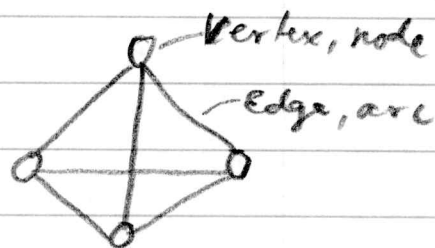
Communities: How things happens in networks.

not guaranteed to find optimal solution
"Find highly interconnected sub-populations"
(Disease, terrorists, network blockage)

heuristic

Basic Graph Theory:

if a set has all edges between each, it's a "CLIQUE"



Louvain Algorithm: Maximize Modularity.

(i) (j) a_{ij} : weight of arc/edge (zero if no edge) between i and j

W_i : ~~the~~ Height of edges connected to node i

W : weight of all edges. ($\sum a_{ij}$)

$$\text{Modularity} = \frac{1}{2W} \sum_{i,j \text{ in same community}} \left(a_{ij} - \frac{W_i W_j}{2W} \right)$$

if min instead of max,

each node will remain its own community

Step 0: Each node is its own community

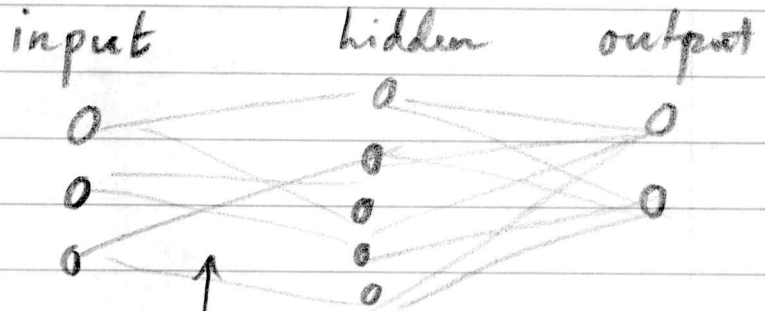
Step 1: Does moving node i to node j 's community maximize modularity?
 Do it

Step 2: Each community is a supernode
 Until no change in mod.

repeat with supernodes.

Neural Networks : Respond to pattern
we don't recognize.

3 "neurons" (basic)



Deep learning

!!! but

lots of layers

many possible layers.

eg. with
gradient
descent

- Send weighted value to next layer
- Pick whichever output has highest result
- Readjust weights
- Keep doing until ~~weights~~ weights don't change

They kinda suck

- Need lots of data to train
- Tough to tune learning algo
- So weights don't change slow or fast

Randomizing, Competitive Decision Making

So far, system does not react. But what if it does? eg: Company sets price, competitors change in response.

Basically, Game Theory.

"Co-operative Game Theory": Competition AND Cooperation.

- Strategies — Pure: make only one choice

Mixed/Random like in Rock, paper, scissors

- Information — Perfect: know ^{anything} about the opponent's situation

Imperfect: Don't know, or have more info than opponent.

- Sum — Zero, when Total "benefit" is zero
One win, one loss.

Non-zero: Total "benefit" is higher or lower or zero

Use optimization models for best strategies

OPTIMIZATION

This only solves stuff. We still build models.

Three Key Components

VARIABLES : The decisions to make.

CONSTRAINTS : Restrictions on variables

OBJECTIVE FUNC : Measure the quality of the solution;

Of a set of values.
Max or min

Have to be explicit about this

Can include inputs from other models!

→ lead to → SOLUTION (Values of variables)

Feasible

Satisfies constraints

Optimal

✓ Feasible soln.
With best objective value.

Binary Variables: There are awesome and allow for more complex models.

- You ^BVar
- ① Define in VARIABLES
 - ② Add to objective function
 - ③ Link variables in constraints!

Examples are common-sensical!

Constraint

"Invest in Tesla" $y_{\text{Tesla}} = 1$

Variables
 $y_i = \begin{cases} 1 & \text{invest} \\ 0 & \text{no} \end{cases}$

"Invest in A, B, C" $y_A + y_B + y_C \geq 1$

"Invest in both A, B or none" $y_A = y_B$

"Invest in A but not B, or vv" $y_A = (1 - y_B)$

"Invest in at least 5 of A" $\sum_{j \in A} y_j \geq 5$

Obj Func

add something

like

$$\sum_i t_i y_i$$

→ Eg: Fixed transaction fee for each y_i (could be fixed)

Op: Linear Regression

$$\min [y_i - (a_0 + \sum_{j=1}^m a_j x_{ij})^2]$$

OPTIMIZE THIS

Statistical Model

Var coeff
 x_{ij} a_j

Optimization Model

Var coef
 a_j x_{ij}

Lasso regression

$$\sum_{j=1}^m |a_j| \leq T$$

Ridge Regression

$$\sum_{j=1}^m a_j^2 \leq T$$

Elastic Net



See Lectures for Stuff

Classification of Models

General : x vector of vars

Min or Max $f(x)$

$s.t. x \in X$ ← Constraint

EASY
FAST

LINEAR : $f(x)$ is linear.

Constraint set X is defined by
linear eqs.

$$\sum_{i=1}^n a_{ij} x_i \leq b_j \text{ or } \geq b_j \text{ or } = b_j$$

not as quick

for each constraint j

Convex Quadratic : $f(x)$ is convex, quadratic

but! Constraint set X is defined
by linear eqs.

Convex Program : $f(x)$ is concave (Maximizing)
or convex (minimizing)

and X is convex

HARD

Integer Program: Linear but!

X should contain some or all integers....

General Non-Convex: prob that is just not convex...

Easy/Quick

Linear

Convex Quad

Convex

Integer

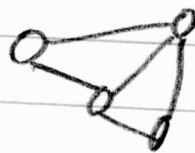
Hard/slow

General non-convex

If too hard, use heuristic!

Network Model

x_{ij} : flow from $i \rightarrow j$



Constraints: - flow into and out of node
- flow across edge

Obj Function: linear function of vars.

- Shortest Path model (GMaps)
- Assignment Model (workers $\rightarrow$ jobs)
- Max. flow model (oil through pipes)

Uncertainty in Optimization

→ Thinks all data is known

- ① Model Conservatively. Add param to constraints (eg. Extra workers just in case)
(demand) $d_j + 0 \epsilon$
- ② Scenario modeling.



"Robust Soln"

Satisfies all scenarios.

S1 shit happens

S2 " "

S3 " "

S4 " "

d_i — demand on day i for scenarios

Shit will get hard, quick.

Optimization models are Math Prog. Models

Var
↳ const
↳ obj func

But can do ~~by~~ Dynamic Programming

- Dynamic Program —
 - States decisions
 - use Bellman's Equations to optimize.
 - add prob. of state transitions
- Stochastic Dynamic Program
 - (w/ discrete states and decisions)
- Markov Decision process
 - (prob. depend only on current state and decision)