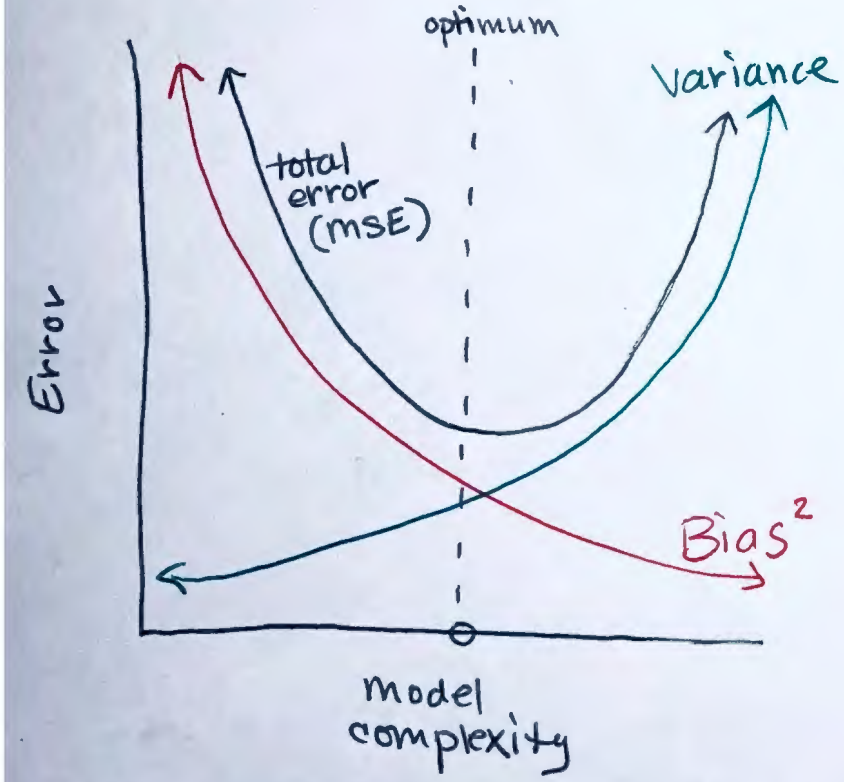
Bias – Variance Tradeoff



Bias = error from missing relevant relations btwn features & output

e.g. underfitting

- High in: linear algorithms
  resampling (model selection)
  simpler models
- The Fix: increase model complexity, reduce regularization

Variance = error from sensitivity to small fluctuations (noise) in training data

e.g. overfitting

- High in: decision trees
  SVM
  KNN
- The Fix: bagging, bootstrapping, regularization, more traindata

Regularization

Def: a process of adding additional info in order to solve a problem

→ adding constraints or penalty to model

Why use it?

- reduce overfitting + increase generalization

How it works

minimize (Loss (Data|model) + λ complexity (model))

weights of features

e.g. L_2 Regularization term = $\sum (weight_n)^2$

What methods?

- Ridge + LASSO regression

Correlation

Def. direction & strength of linear relationship between two variables

value lies between -1 and 1

Methods

Pearson's R

$$r = \frac{\text{covariance}(x,y)}{\text{stdv}(x) \cdot \text{stdv}(y)}$$

Spearman's Rank

(one var increases as the other var does)

- assesses monotonic relationships

Point-Biserial

- assesses relationship between continuous & binary vars

! **Note:** Spurious correlation are things with a mathematical relationship, but no real-life relationship

XGBoost

➡ accuracy, efficiency, feasibility

Def. a linear model & tree learning algo that does parallel computations on a single machine; a fast implementation of gradient boosted trees

Ridge Regression

L2 Regularization method that shrinks the size of the coefficients

Encourages:

- coefficients that are closer to true population parameters
- equal shrinkage

Combats:

- overfitting
- multicollinearity
- sparseness (avoids it)
- high variance

Other Uses:

Cons:

- sensitive to outliers

R code:

```
> library(glmnet)
> model = cv.glmnet(
      x, y, alpha = 0)
```

minimize $\rightarrow$ $\sum_{i=1}^{n} (y_i - \sum_{j} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$

$\hat{y}_i$

$\sum_{i=1}^{n} (y_i - \sum_{j} x_{ij} \beta_j)^2$: sum of squared error

λ: tuning param. of strength of penalty

$\sum_{j=1}^{p} \beta_j^2$: penalty

y_i = Response data point x_{ij} = covariate matrix β = coefficients

LASSO Regression

a.k.a. Least Absolute Shrinkage & Selection Operator

L1 Regularization method that shrinks data points towards a central value, like the mean

Encourages:

- simple, sparse model
- fewer parameters by elimination
- shrinking β to exactly $\emptyset$

Combats:

- overfitting
- multicollinearity
- high variance

Other Uses:

- automate feature selection

R code:

```
> library(glmnet)
> model = cv.glmnet(
      x, y, alpha=1)
```

minimize →

$$\underbrace{\sum_{i=1}^{n}\left(y_i - \overbrace{\sum_{j} x_{ij}\beta_j}^{\hat{y}_i}\right)^2}_{\text{sum of squared errors}} + \underbrace{\lambda}_{\substack{\text{tuning parameter} \\ \text{of strength} \\ \text{of penalty}}} \underbrace{\sum_{j=1}^{p}|\beta_j|}_{\text{penalty}}$$

y_i = response var data point x_{ij} = covariate matrix β = coefficients

Bootstrapping

Def: a resampling method that uses sampling with replacement to simulate multiple samples/experiments

Uses

- estimate C.I.
- estimate standard errors

Benefits

- non-parametric (e.g. distribution-free)
- recreates any # of resamples

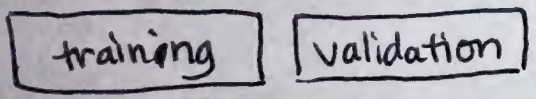
R Code

```
> library(boot)
> boot(data, statistic, R= # of resamples)
```

Cross-Validation

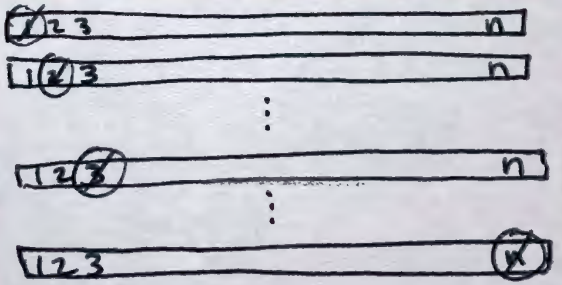
Def: a resampling method used to evaluate a model's performance (generalization)

Validation Set



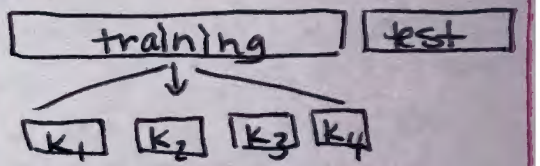
- simplist approach
- can overestimate accuracy

Leave One Out CV

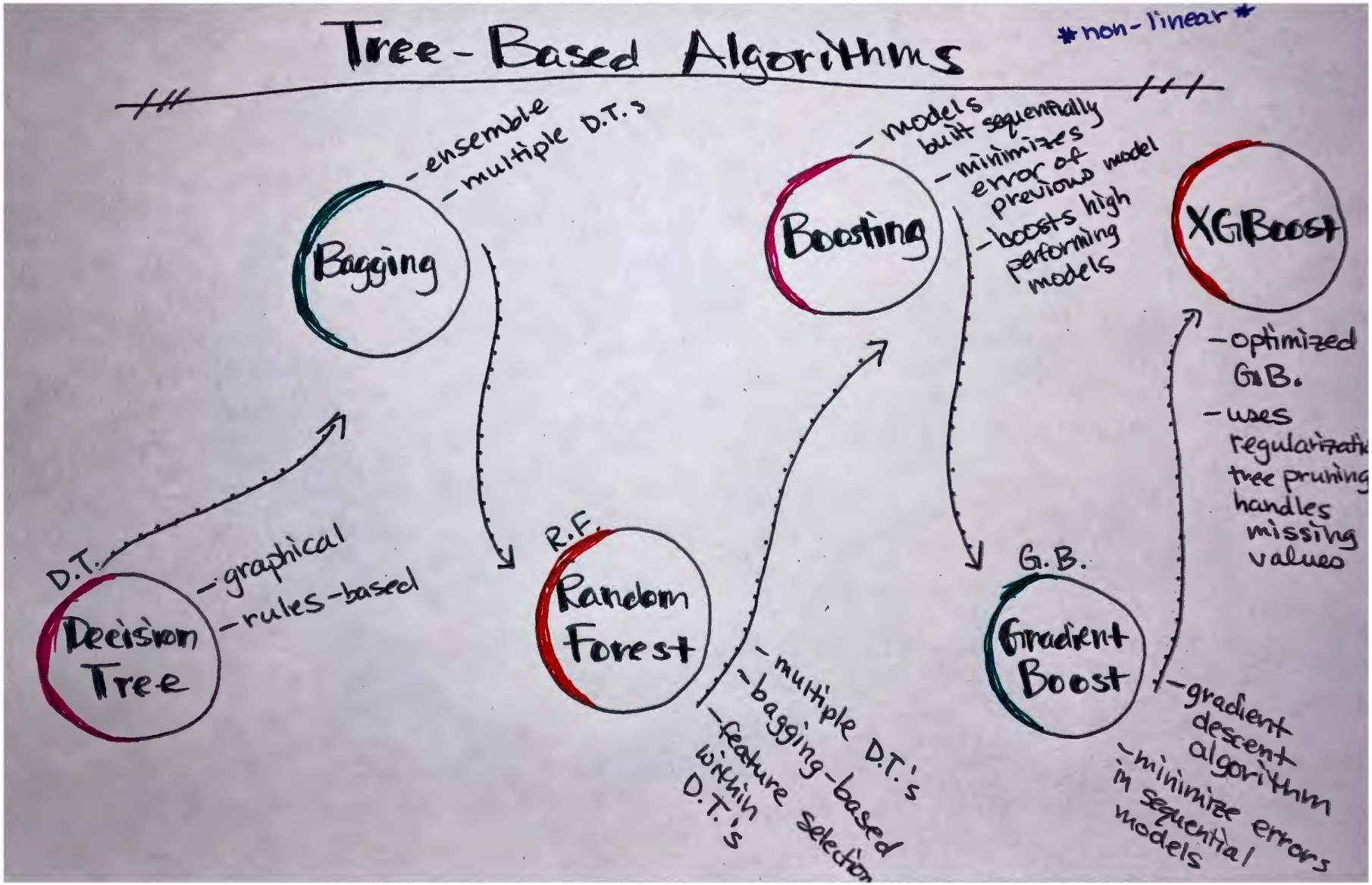


- highly unbiased
- high variance
- computationally expensive

K-fold CV



- typical K = 5 or 10
- middle of the road bias + variance
- not as computationally expensive


Tree-Based Algorithms
non-linear
Bagging
- ensemble
- multiple D.T.s
Boosting
- models built sequentially
- minimizes error of previous model
- boosts high performing models
XGBoost
- optimized G.B.
- uses regularization tree pruning handles missing values
D.T.
Decision Tree
- graphical
- rules-based
R.F.
Random Forest
- multiple D.T.'s
- bagging-based
- feature selection within D.T.'s
G.B.
Gradient Boost
- gradient descent algorithm
- minimize errors in sequential models

Type I & II Errors

"false positive"

Type 1 Error

- mistaken rejection of an actually true null hypothesis

a.k.a saying 2 groups are different when they are not

Tolerance for this error set by your **alpha α level**

confidence level = $1-\alpha$

e.g. $\alpha = 5\%$, 5% chance of a Type 1 error

"false negative"

Type 2 Error

- mistaken acceptance of an actually false null hypothesis

a.k.a saying 2 groups are not different when they are

Tolerance for this error set by your **power level**

e.g. power = 80%, 20% chance of Type 2 error

Linear Regression *supervised*

Def: a method of fitting a linear trend line to data points in order to predict future values

Goal minimize the sum of squares of the difference between the observed ~~data~~ & predictions ← variance

Assumptions

linear relationship; features are not correlated; errors are uncorrelated & homoscedastic & normal

Validation

$$y_i = \beta_0 + \beta_1 X_i \ldots + \varepsilon$$

R^2 = goodness of model fit; explanatory power of features

adj. R^2 = same as R^2 but penalizes for more features

AIC = used for model selection; estimator of prediction error

Residual plots & Durbon-Watson = check if residual assumptions are met

Naive Bayes

supervised

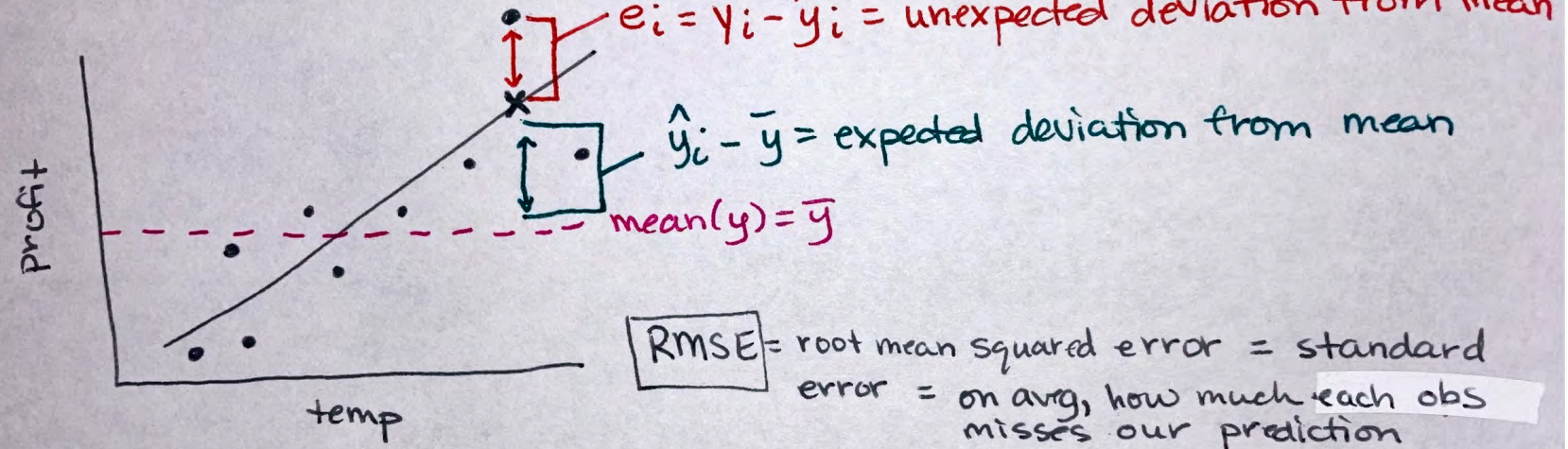
Def: probabilistic classifier based on Bayes' Theorem with strong independence assumptions between features

- not iterative, so very scalable
- uses MLE
- used often for text classification + recommendation systems

$$P(A \mid x_1, \ldots, x_n) = P(x_1 \mid A) \cdot P(x_2 \mid A) \cdot P(x_3 \mid A) \ldots \cdot P(x_i \mid A) \cdot P(A)$$

Error

pertains to how well a model explains the true population



RMSE = root mean squared error = standard error = on avg, how much each obs misses our prediction

SSR = sum squared regression = $\sum(\hat{y}_i - \bar{y})^2$, also called ESS

SSE = sum squared error = $\sum e_i^2$, also called RSS

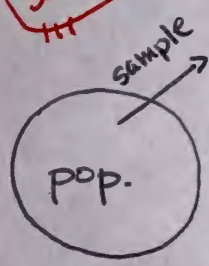
SST = SSR + SSE = sum squared total, also called TSS

MS = mean square = sum squared errors / degrees freedom = unbiased estimate of error variance

↑ use this to calculate F-statistic to accept/reject Null that all regression coefficients equal zero

C.I. - Proportion

one sample



$\hat{p}$ = sample proportion

Conditions for Inference

- random
- normal $n \cdot \hat{p} \geq 10$, $n \cdot (1-\hat{p}) \geq 10$
- Independence

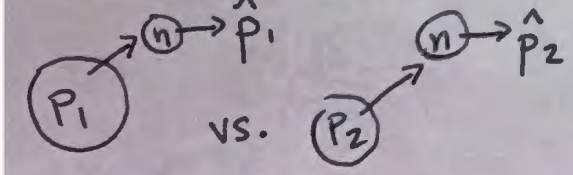
Steps:

1) choose cofidence level, e.g. 95%
2) obtain critical value for that confidence level, e.g. z-score
3) calculate C.I.

$$\hat{p} \pm z\text{-score}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$\sigma_{\hat{p}}$ ← if you know the true population stdv., use that

two samples



confidence interval for $(p_1 - p_2) = (\hat{p}_1 - \hat{p}_2) \pm z\left(\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$

standard error → $\sigma_{\hat{p}_1 - \hat{p}_2}$

unsupervised

K-Means

hard clustering

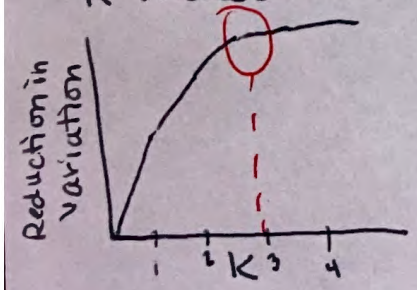
Def: unsupervised clustering algorithm that groups data points based on the nearest mean (cluster center)

Goal: minimizes within-cluster variance (squared Euclidean distance)

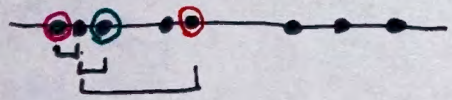
Steps:

① Define # of clusters

- use elbow method to plot total variation within cluster for diff. K values

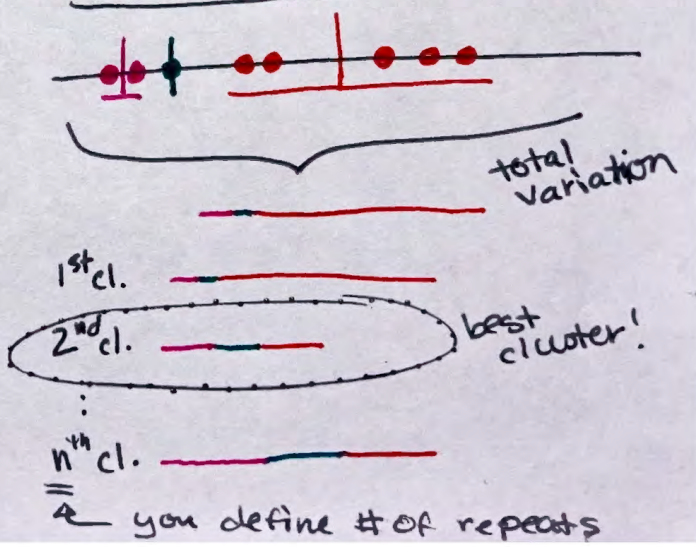


② Randomly select K datapoints to be initial clusters



- assign rest of points to cluster based on shortest distance
- after assigned, calculate mean of each cluster + re-assign based on point distance to cluster mean

③ Repeat clustering method (#2) until best total variation within cluster is reached



unsupervised*

Mixture Models + EM Algo

soft clustering

Def: unsupervised clustering method based on probability distributions (Gaussian or multinomial)

Goal: discover unknown mean & variance of distributions

Steps:

(1) Start by placing the K distributions somewhere

- find the (μ, σ^2) for each one
- calculate probability that data point belongs to each distro
- softly assigns point to group based on likelihood to be in that group

(2) After first probabilities found, adjust distro (μ, σ^2) to fit points & iterate process

- maximize the log-likelihood function

↑ the most probable assignments

Hierarchical Clustering

unsupervised

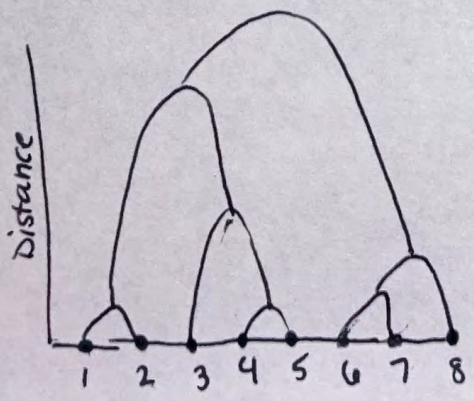
Def: an algorithm that builds a hierarchy of clusters based on a defined distance metric

e.g. Euclidean distance, squared Euclidean distance, Manhattan distance, maximum distance, mahalanobis distance

Agglomerative

"bottom-up"

- each obs. starts as its own cluster



Divisive

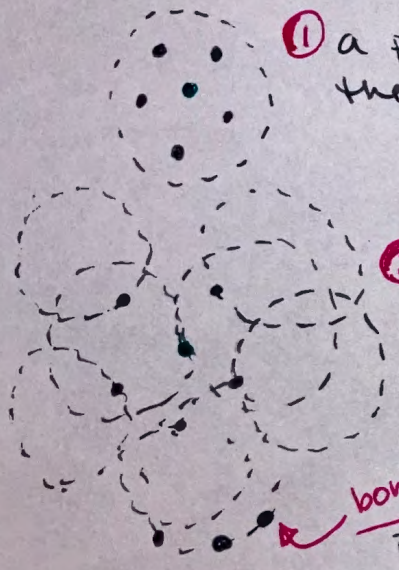
"top-down"

- all obs. start in 1 cluster

DBSCAN

unsupervised

Def: clustering algorithm that groups any shape



1. a point is randomly selected as starting point & based on the "minimum points" and the declared epsilon (distance)

2. each point in the initial cluster will broadcast out their perimeter, looking for new point members

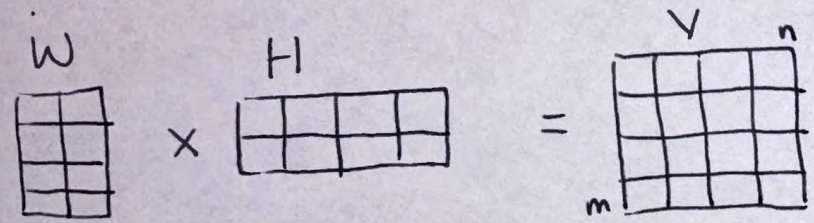
border points
part of cluster, but not within epsilon

3. border points broadcast out their perimeter to gain new members

4. method is repeated for any remaining, unassigned points

Non-Negative Matrix Factorization

Def: a dimension reduction algorithm which decomposes a matrix V into 2 lower dimension matrices



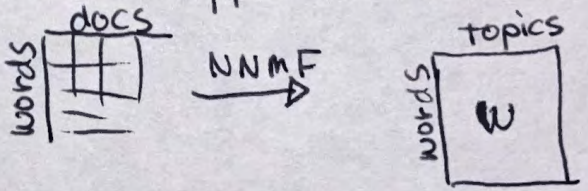
- each element in matrix ≥ 0

each column is a "basis" element, or a component that comes up many times in the data points

tells you how to reconstruct the original data point, given the basis elements

Cool!

when applied to text & documents



W matrix colums represent topics

Multicollinearity

Def: the state of independent variables being correlated to each other in regression ; not a good thing ; violation of assumptions

Why is it bad?

- muddies interpretation of marginal effects of each variable, since "holding other variables constant" does not apply
- difficult to determine which vars actually had an effect on Y
- inflates variance of affected variables → increases standard errors & p-values

Detection

- correlation of independent variables
- VIF

The Fix

- Do nothing!
  - ok if model is only being used for prediction (doesn't affect model fit)
- remove correlated vars
- combine vars or use PCA

Applicable to

- linear regression
- generalized linear model
- logistic regression
- Cox regression

Margin of Error

confidence interval: $\mu \pm z\text{-score} \cdot se$

margin of error: $z\text{-score} \cdot se$

mean

$$ME = z\text{-score} * \frac{\sqrt{s^2}}{\sqrt{n}}$$

proportion

$$ME = z\text{-score} * \sqrt{\frac{p(1-p)}{n}}$$

Precision + Recall

Def: classification metrics

Accuracy tells us $\frac{\text{total correct}}{\text{total obs.}}$, but can have issues with imbalanced classes

Precision

$$= \frac{\#\text{ true positives}}{(\#\text{ true positives} + \#\text{ false positives})}$$

e.g. $\frac{\text{correct results}}{\text{all returned results}}$

Want high for → when there's a high cost of false positives (pregnancy test, recommendation)

Recall

$$= \frac{\#\text{ true positives}}{(\#\text{ false negatives} + \#\text{ true positives})}$$

e.g. $\frac{\text{correct results}}{\text{results that could have been returned}}$

Want high for → when there's high cost of false negatives (identifying fraud or cancer)

Expected Value & Variance

Discrete

$$E(x) = \sum x_i \cdot p(x_i)$$

- weighted average
- for Binomial = $n \cdot P$

$$\sigma^2 = E[(X-\mu)^2]$$

$$= \sum((x_i-\mu)^2 \cdot p(x_i))$$

- for Binomial = $n \cdot P \cdot (1-P)$

Continuous

$$E(x) = \frac{\sum x_i}{n}$$

$$\sigma^2 = \frac{\sum (x_i-\mu)^2}{n-1}$$

Correlated Topic Modeling

ef: a probabilistic approach to infer latent topics of a document ; allows latent topics to be correlated

pplication: derive themes or topics within book (or corpus of books)

sumptions:

each document is comprised of multiple topics, made of multiple word
order of word doesn't matter
word distribution per topic & topic distro per document is logistic norma

Word Network Topic Modeling

Def: method to identify topics within short text using network graphs

Method

1) create word network of document text & apply an edge weighting of occurance frequency of word pairs
2) apply LDA to the graph's adjacency matrix
3) LDA gives you topics for each word & will need to be mapped back to documents

tf-idf

" term frequency - inverse document frequency "

Def: measure of how important a word is to a document in a collection (corpus)

→ $idf(term) = \ln\left(\frac{n_{documents}}{n_{documents\ containing\ term}}\right)$

→ $tf(term) = n_{occurances\ of\ term}$

$$tf\text{-}idf(term) = tf \times idf$$

when low....

e.g. tf-idf = 0

extremely common word

when high....

e.g. tf-idf = 0.009

important words

P-Value → probability of observed outcome

Def: the probability of obtaining test results at least as extreme as the results actually observed

→ used in hypothesis testing

....Calculation.....

T, test statistic, provides a single number & follows a distribution determined by the function used to define the test statistic & distro of input data (observations)

e.g. Z-tests → t-statistic ; normal distro w/ known variance

t-tests → t-statistic; normal distro w/ unknown var.

F-tests → F-statistic; for variance

Chi-squared test → chi-squared-statistic ; discrete data

To calculate a p-value, you need ① null hypothesis, ② test statistic (1 or 2 tailed), ③ data

the Sign Test

non-parametric

Def: statistical method to test for consistent differences between pairs of obs.

e.g. weight pre-treatment (x)
weight post-treatment (y)

When to use

when determining if ... $x > y$ $x = y$ $x < y$ $\text{median}(x) \gtrless$ some #

Assumptions

- data randomly selected
- samples must be dependent or paired

! will not work on 'independent samples'

Assumption of Normality

Inferential statistics rely on the assumption that the value of interest will exhibit a bell-curve distribution if many random samples are taken. → parametric

* means across samples assumed to be normal, not that the sample obs. or population obs. are normal

also called sampling distro

Why do we assume a normal distro? → Central Limit Theorem

Methods that rely on assumption

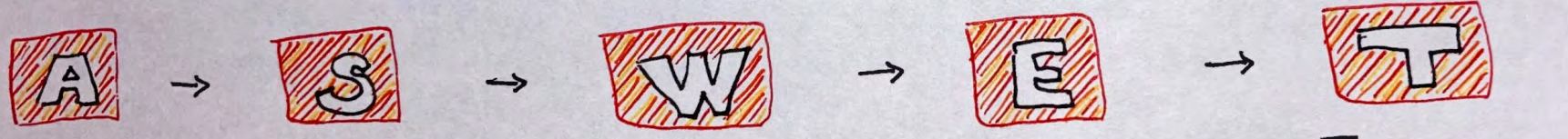
t-tests, ANOVA, regression

Requirements to use this assumption

N sample size needs to be large enough

Data Science in Business

1. What business question are you solving?
2. What methods will you use?
3. How will you evaluate performance of methods?
4. How do you generalize it to similar business problems?



Ask Questions

- assumptions
- historical data
- KPI
- define vagueness

Solution

- start simple

Explain Why

- reason for approach

Evaluate Solution

- technical evaluation
- business evaluation

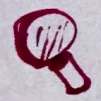
Trade offs

- pros/cons

Endogeneity

Def. when an explanatory variable is correlated with the error term because of measurement error or confounding in omitted or unobserved variables

relevant for: time series, causal analysis

ow to Detect it 

How to Correct It ✓

- Instrumental variables

Standard Error

Def: the measure of uncertainty in the sample mean

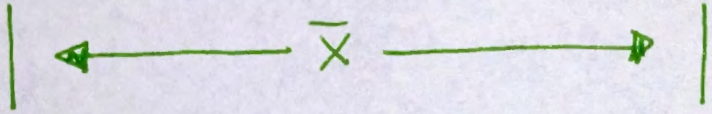
e.g. how precise is our estimate of the mean?

* decreases as sample size grows

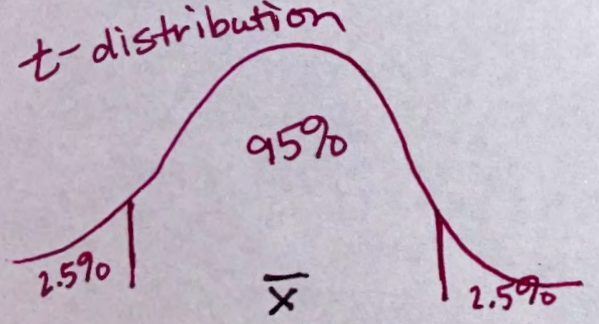
sample sd

sample mean	sample proportion	diff. of means	diff. of proportions
$SE(\bar{x}) = \frac{s}{\sqrt{n}}$	$SE(p) = \sqrt{\frac{p(1-p)}{n}}$	$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$SE(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Confidence Interval



Def: an interval in which the population mean lies within, with a certain % of confidence



$$\bar{x} \pm SE(\bar{x}) * t\text{-statistic}_{0.975,\, n-1}$$

* both part of normal distro

Z score	vs.	T-score
• use when $n > 30$ or population is known		• use when $n < 30$

7 Common Distributions

Name	f(x)	E(x)	Var(x)	Use Case
Bernoulli	p x=success $1-p$ x=failure	p	$p(1-p)$	single coin toss
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	$n \cdot p$	$n \cdot p \cdot (1-p)$	n Bernoulli trials
Geometric	$p(1-p)^x$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	# of failures before first success
Uniform (continuous)	$\frac{1}{b-a}$, $x \in [a,b]$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$	random # generator
Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2	
Poisson	$\frac{\lambda^x}{x!} e^{-\lambda}$	λ	λ	# of phone calls in one day *related to binomial
Exponential	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$(\frac{1}{\lambda})^2$	how long it takes between 2 calls *related to poisson

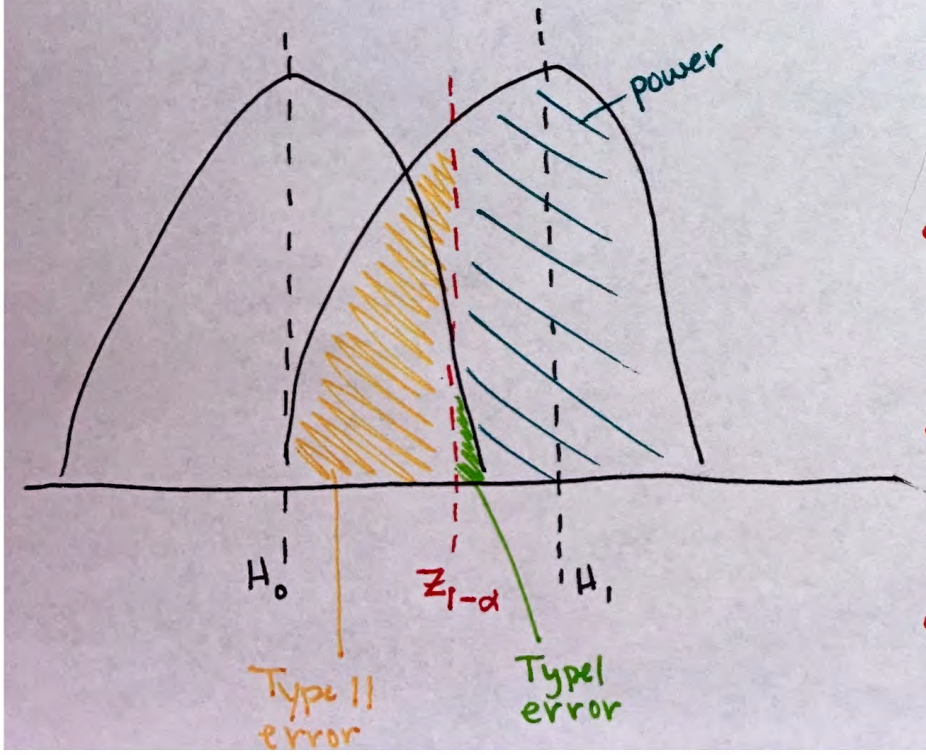
Impacts of Sample Size

SE = standard error (dispersion of sample mean distribution)

power = $1-\beta$ (probability of not getting Type II error)

confidence level = $1-\alpha$ (probability of not getting Type I error)

effect = how big a difference is practically significant (minimum detectable effect, MDE)



- as sample size ↑, SE ↓ because sample mean gets closer to population mean (less dispersion)
- as sample size ↑, power ↑ bc the range of acceptance decreases (smaller "area" of TII errors)
- as sample size ↑, confidence ↑ bc the area of the distribution included within the acceptable range increases
- as sample size ↑, MDE ↓

Sample Size

Def: total # of participants that adequately represent the population

Assumptions:

- every individual has equal chance of being put in sample
- participants don't affect each other

$$n = \frac{2\left(z\text{-statistic}_{\alpha} + z\text{-statistic}_{1-\beta}\right)^2 \cdot s^2}{\text{estimated effect size}}$$

n = required sample size

Types of Hypothesis Tests

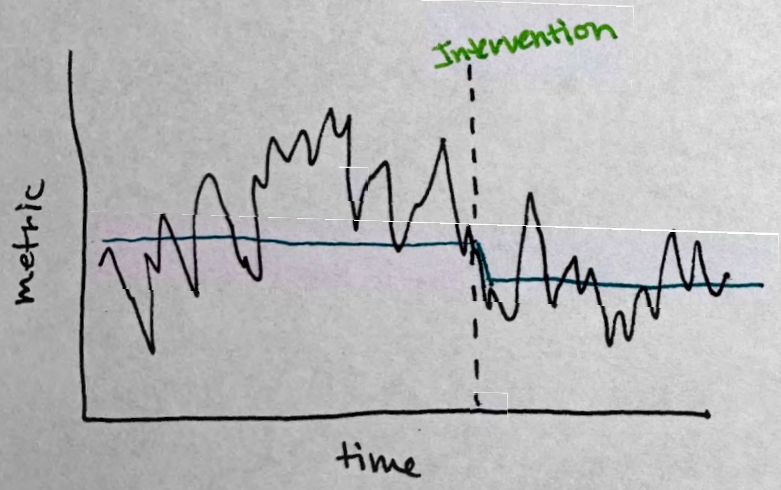
Name	When to Use	Test Statistic	Underlying Distro
t-test	• unknown mean & variance • small samples • finding means, difference	$t = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$	t-distribution * at infinity this becomes Normal
z-test	• unknown mean, known variance • large samples • finding means, proportions, differences	$z = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$	normal
chi-squared	• tests independence of a pair of variables • tests if there's a relationship (no direction) • categorical data (contigency table)	$\chi^2 = \sum \frac{(\text{obs. value}_i - \text{expected value}_i)^2}{\text{expected value}_i}$	chi-squared
Fisher's exact test	• tests independence • alt. to chi-squared when n is small • categorical data (contingency table)	$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$	hypergeometric
Mann-Whitney U-test	• nonparametric • lower power • data from both groups need to be symmetric & identical • specific Null hypoth.		
ANOVA (F-test)	• tests mean between 3+ groups • tests explained variance vs. unexplained in model	$F = \frac{\text{between group variability}}{\text{within-group variability}}$	F-distribution
Welch's t-test			

Gathering Business Requirements

- What customer problem are we trying to solve?
- How will we know we've completed what's required?
- What does success look like?
- How will we measure success?
  - What specific metrics indicate success vs. failure?
  - How will those metrics be measured?
  - Who is responsible for taking those measurements?
  - When will the measurement take place?
  - Who needs to review the measurement & make the call?
- What decision will be made based on the outcome?
  - Who needs to approve that decision?
- What is the business question or purpose?

Intervention Analysis

Def: a time series analysis that estimates the effect of an external intervention on a time series, assuming the same ARIMA structure holds before + after intervention



The effects of the intervention are evaluated by changes in the level + slope of the time series + statistical significance of the metric

Time Series

Def: analysis of data that changes over time; using a variable's own past values to forecast its trend

Types of Models

Autoregressive (AR) = current data is built upon from data at previous time; how far back do we need to look back?; e.g. stock prices

Moving Average (MA) = current data is a result of previous unexpected events; how many days of unexpected events do we look at?

Autoregressive Moving Average (ARMA) = combination of AR + MA models

Assumptions:

- constant mean + constant variance for whole series → Stationary
  - use Lag Differencing to transform non-stationary data
- errors can be written as a linear function of past observations → Invertible