(closed book, simple calculator)

Lecture 2: Regression Analysis

Confidence Interval

Provide us w/ range of values that we believe, with % level of confidence, contain a population parameter

$$Pr\left(\bar{x} - z_{(1-\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{(1-\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}} \right)$$

α = % level confidence

Larger sample = smaller interval

Lower α = smaller interval, larger variance = larger interval

use z if know population σ, use t if know sample s

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

degree of freedom $df = n-1$ → use in the t-dist. graph

Hypothesis Testing

1. State null hypothesis H_0 and Alternative hypo. H_1
2. Choose α - level

or Calculate p-value

if $p \leq \alpha$ reject H_0

$p > \alpha$ cannot reject H_0 (not accept H_0 this)

H_0	H_1	Reject H_0 if	
$\mu = \mu_0$	$\mu \neq \mu_1$	$\lvert z \rvert \geq z_{1-\alpha/2}$ or $\lvert t \rvert \geq t_{1-\frac{\alpha}{2},n-1}$	(2 sided test)
$\mu \leq \mu_0$	$\mu > \mu_0$	$z \geq z_{1-\alpha}$ or $t \geq t_{1-\alpha,n-1}$	(one sided test)
$\mu \geq \mu_0$	$\mu < \mu_0$	$z \leq z_\alpha = 1 - z_{1-\alpha}$ or $t \leq t_{\alpha,n-1}$	(one sided test)

Type I Error : False Positive - Reject H_0 when H_0 is actually true

Type II Error : False Negative - Fail to reject H_0 when H_1 is true

P-Value and t-Test explanation

The hypothesis that there's no difference between 2 measured is called Null Hypothesis (H0). In T-Test, we check with 95% (more or less) confidence interval which is the middle curve part of the normal distribution.

If the **t test is higher** than the t value of H0 at 95% confident. Then that mean that the new distribution is not inside the 95% of H0 and there's enough difference between the 2 distribution, so we can **reject the H0** (it does not exist within H0 area).
If **it's lower**, then it's inside the 95% H0 area and the difference is not significant enough to reject the H0 (that there's no difference) so we **fail to reject** H0.

P-Value work similarly. If p-value is large, then the difference between 2 distribution is "not significant" enough, meaning there's no significant diff between default and the one we're measuring. The measured data is close enough to the distribution. If we create the distribution on the measured data, it will be not be significantly difference from the first distribution so there's high probability that we can observe the value under H0 area. (and fail to reject H0)
If the p-value is small, then the difference between 2 distribution is "significant", meaning the measured data is far away from the distribution and other distribution would do a better job explaining the data. If we create distribution on measured data, the distribution will be significant away from the 1st distribution and there's low probability that we will observe the value under H0 area (and reject the H0)

Alternative Hypothesis is **another hypothesis that is true if null hypothesis is false**. It's mostly the opposite of null hypothesis (In this case, there's a difference between 2 measured)

t-Test Type

Single Sample - test whether $\bar{x}$ is significantly diff from a pre-existing value

Paired sample - test relationship btw 2 linked sample eg. mean obtain in 2 condition by a single group

Independent sample - test relationship btw 2 indep. problem

Paired t-Test w/ 2 Paired Sample

$$H_0 : \mu_d = \mu_1 - \mu_2 = D_0$$

Test statistic $$t = \frac{\bar{d} - D_0}{\frac{s}{\sqrt{n}}}$$ where $\bar{d}$ = mean of the diff btw 2 sample test

$H_1 : \mu_d \neq D_0$ Reject H_0 if : $|t| \geq t_{1-\frac{\alpha}{2}, n-1}$

$: \mu_d > D_0$ $: t \geq t_{1-\alpha, n-1}$

$: \mu_d < D_0$ $: t \leq t_{\alpha, n-1}$

Independent Sample – 2 indep. sample

$H_0: m_1 = m_2$

$H_1: m_1 \neq m_2$

Welch's t-Test

2-sample unpaired t-test w/ equal or not equal sample size, assuming unequal variance

$df = n_1 + n_2 - 2$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

where $s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Linear Regression

Relationship btw dependent variable (y = label/outcome) and indep. variable (x = all the input feature)

(response, predictor)

Simple Linear Regression model

$Y = \beta_0 + \beta_1 X + \varepsilon$ — β_0, β_1 are unknown coefficient

β_0 = y-axis intercept β_1 = slope of the line ε = random error term

Ordinary Least Square (OLS) Estimators

Which line fit data the best? → line that minimize Sum of Squared Residual (loss func)

let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ $RSS = \min \sum_i^2 (y_i - \hat{y}_i) = \sum_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x))^2$

then $\hat{\beta}_1 = \frac{cov(x,y)}{var(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$ $\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Note

$cov(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

$Var(x) = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$

Residual Sum of Square

$RSS = \sum_i (y_i - \hat{y}_i)^2$

$ESS = \sum_i (\hat{y}_i - \bar{y})^2$ → Explain sum of squared

$TSS = \sum_i (y - \bar{y})^2 = ESS + RSS$ = Total sum of square

R^2 measure proportion of variance in y that is explain by variance in X

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} \quad (TSS = ESS + RSS)$$

$R^2 = 1$ – perfect match btw line and data

$R^2 = 0$ – no linear relationship btw x & y

Multiple Linear Regression – $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_n x_n + \varepsilon$ $\quad Y = X\hat{\beta} + e$

OLS estimator – min error = $\min RSS = \min (Y - X\hat{\beta})^T (Y - X\hat{\beta}) = \min e^T e$

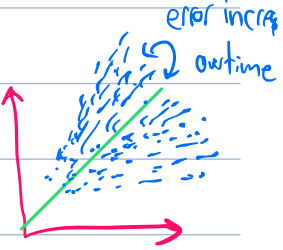
$\frac{\partial RSS}{\partial \beta} = 0 \rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$ and so $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$ (normal equation)

Lecture 3 : Regression Diagnostics (BLUE)

Gauss-Markov Theorem

the OLS estimator is the best linear unbiased estimator (BLUE), iff the model has

1. Linearity - Linear relationship in parameters β
2. No multicollinearity of predictors → $Cov(x_i, x_j) = 0 \ \forall i \neq j$
   No linear dependency btw predictors
3. Homoscedasticity - The residual (error) exhibit constant variance
   violate if variance increase overtime



4. No autocorrelation - no correlation btw i and j residual term (error)
5. Expected value of the residual vector (error) given X is 0 $E(\varepsilon|X) = 0$
   exogeneity $cov(\varepsilon, X) = 0$

unbiased = $E(\hat{\beta}_j) = \beta_j$

best estimator = give lowest variance compare to other unbiased estimators (some biased can be better)

1. Linearity

For non linear → reformulate model that make it linear again

- Polynomial regression - $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \varepsilon$ let $z = x^2 \rightarrow Y = \beta_0 + \beta_1 X + \beta_2 z$ → linear now
- Transform X and/or Y - $\log(Y) = \beta_0 + \beta_1 \log(X) + \varepsilon$
- Split data $X[X > X_k]$ (choose part of data that is linear)

2. No multicollinearity of predictors

Indep. var must not be linearly dependent. If 2 indep. var. are dependent → omit one of them

- Check the rank of the matrix X must have full column rank $rank(X) = p$ p = # of col of X

High correlation btw indep. var leads to issue wrt. significance of predictors

Check for multicollinearity

Calculate the correlation coefficient btw every variable (both indep. and dep. var)

Correlation = ±1 → high correlation Correlation = 0 no correlation

Variance inflation factor (VIF) $= \frac{1}{1 - R_k^2}$ where R_k^2 = how much % of variance of predictor in question can be explained by other indep. var

if VIF >10, 90% of var has been explain by other. So we can remove that variable

If variable has low t-value or large p-value (non significant), then either

1. Variable is not relate to response (y) → small t, small VIF, small correlation w/ response
2. Relate to response (y) but strongly relate to other variable → small t, large VIF, large correlation w/ response

(Look at example how variable change → p.20-22

Homoscedasticity

residual has a constant variance → Homoscedasticity

if residual doesn't have constant variance → Heteroscedasticity $Var(\varepsilon_i | x_{1i}, \ldots, x_{pi})$ not constant

Test to check for homoscedasticity – Glejser test, White Test

No Autocorrelation

No pattern should be detect when graphing residual over time

Modelling seasonality – use season instead of time b/c time has autocorrelation but season does not

- Create dummy variable which indicate season
  - → have a time period to indicate which season is the data in
  - → have 3 column for 3 season (Q_1, Q_2, Q_3), 1 if it's in one of the season, 0 if not
  - → if all col is 0, then it's the last season → if we model all 4 season → multicollinearity

$$y = \beta_0 + \beta_1 t + \beta_2 Q_1 + \beta_3 Q_2 + \beta_4 Q_3$$

Expected value of the residual vector, given X, is 0 $E(\varepsilon|X) = 0$

Endogeneity – $corr(\varepsilon_i, x_i) \neq 0 = E(\varepsilon_i | x_i) \neq 0$

Reason for endogeneity – omitted (confounding) variables

if True model : $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_i$

but our model use : $y_i = \beta_0 + \beta_1 x_1 + u_i$ then $u_i = \beta_2 x_2 + e_i$

if x_1 and u_i are correlated and x_2 affects y, this lead to endogeneity

Technique to address endogeneity

Panel Data vs Cross Section Data

(not care about time)

Cross-section data – observe many subjects at the same point of time, or w/o regard to diff in time

Panel Data Set - repeated observation on the same subject which make it possible to overcome omitted variable bias

- Unbalance - some subject haven't been record in some time period (data not full)
- Balance - every subject is record in every time period

Model fixed Effects

Fixed effect - Treat α_i as a constant for each subject (add to the intercept $\beta_0 + \alpha_i$)

- Idea: α represent everything we don't know about the subject (unobserved heterogeneity).

Get a good estimate for other covariate (other variable that is correlate) w/o knowing the unobserved heterogeneity

Have various estimator for fixed effect model $y_i = (\beta_0 + \alpha_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \varepsilon_i$

Lecture 4: Logistic & Poisson Regression

In the OLS regression, predicted prob. of the linear model can be >1 or <0

In Logistic Regression $= P[Y|X]$ — constrains estimated prob. to lie btw 0 and 1

$$P(Y|X) = p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$ } $P(Y=y|X)$ is a non linear func.

β_0 = constant → move curve left/right

β_1 = slope = steepness of curve

Odds and Logit

$$\text{odd} = \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$ odd range from 0 to ∞ (based on $e^{\beta_0+\beta_1 X}$)

Logits → take "ln" on both side $\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$ → linear prob btw 0 and 1

logits is now between $-\infty$ to ∞ (based on $\beta_0+\beta_1 X$) and β_0 = y-intercept, β_1 = slope

Estimating Coefficients of Logistic Regression

Maximum Likelihood estimate (MLE) → Assume $X \sim \text{Ber}(p)$ so $p(y_i) = p^{y_i}(1-p)^{1-y_i}$

Likelihood func $L = \prod_{i=1} p^{y_i}(1-p)^{1-y_i}$

$P(Y_i=1|X) = p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \sigma(\beta_0+\beta_1 X)$, so $L = \prod_{i=1} \sigma(\beta_0+\beta_1 x)^{y_i}(1-\sigma(\beta_0+\beta_1 x))^{1-y_i}$

find β that maximize likelihood

Log Likelihood $LL = \ln(L) = \sum_i y_i \ln p_i + (1-y_i)\ln(1-p_i) \rightarrow \max_\beta \sum_i y_i \ln(\sigma(\beta_0+\beta_1 x)) + (1-y_i)\ln(1-\sigma(\beta_0+\beta_1 x))$

Gradient Ascent - $x^{(k+1)} = x^{(k)} + \alpha \nabla f(x^{(k)})$ gradient → direction of steepest ascent

Use it to maximize LL or minimize NLL

Goodness of Fit

Null model - model with only the intercept (only estimate 1 param)

Fitted model - model w/ p parameters and intercept term, so we have p+1 param

Null Deviance $= -2\ln(L(\text{null}))$ - How much is explained by a null model

Residual Deviance $= -2\ln(L(\text{fitted}))$ - How well the fitted model explain the data → small value = explain well

McFadden R^2

Similar to R^2, but R^2 is for linear regression, McFadden R^2 is for logistic regression

$$R^2_{\text{McFadden}} = 1 - \frac{LL(\text{fitted})}{LL(\text{null})}$$

If the fitted model explain much more than a null model, then $R^2 \to 1$

If fitted model doesn't explain much → $R^2 \to 0$ ($R^2 > 0.2$ is okay)

logit model is an example of **Generalised linear model (glm)**

GLM - general class of linear model made up of 3 component

1. Distribution for modelling Y (Gaussian, Binomial, Poisson)
2. Linear Prediction $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
3. Link function $g(\cdot)$
   - Identity link (use in normal linear regression) — $g(\mu) = \mu = X\beta$
   - Logit Link — $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
   - Log Link — $g(\mu) = \log(\mu)$

Count variable

Count variable → non negative integer

Can be model w/ OLS regression but it can yield negative predicted value and count variable is often skewed

Use Poisson model or Negative binomial

In poisson, coefficient β_i, when interpret need to be put in e^{β_i} → variable x_i decrease/increase rate by e^{β_i} factor

Lecture 5: Naive Bayes

Prob. can be multiply

Naive Bayes – All feature are equally important and independent of each other

Conditional Prob and Bayes Rule

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(A|B)P(B) = P(B|A)P(A) \quad \text{(product rule)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{(Bayes Rule)} \qquad P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(B) = P(B|A)P(A) + P(B|\neg A)P(\neg A) = P(B \cap A) + P(B \cap \neg A)$$

R = Result of test D = disease

$P(R=P \mid D=P) = 0.98 \rightarrow P(R=\neg P \mid D=P) = 0.02$

$P(R=\neg P \mid D=\neg P) = 0.97 \rightarrow P(R=P \mid D=\neg P) = 0.03$

$P(D=P) = 0.008 \quad P(D=\neg P) = 0.992$

$$P(D=P \mid R=P) = \frac{P(R=P \mid D=P)\,P(D=P)}{P(R=P)} \rightarrow \frac{P(R=P \mid D=P)\,P(D=P)}{P(R=P \mid D=P)P(D=P) + P(R=P \mid D=\neg P)P(D=\neg P)}$$

$$= \frac{(0.98)(0.008)}{(0.98)(0.008)+(0.03)(0.992)} = 0.21$$

$P(D=\neg P \mid R=P) = 1 - 0.21 = 0.79$

Frequency Table

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ → P(Play | Outlook) → what is the prob of play given each outlook

Outlook	Play = No	Play = Yes
Sunny	3	2
Overcast	0	4
Rainy	2	3

Bayes Prob

$p(O=S \mid P=N) = \frac{3}{5}$	$p(O=S \mid P=Y) = \frac{2}{9}$
$p(O=O \mid P=N) = 0$	$p(O=O \mid P=Y) = \frac{4}{9}$
$p(O=R \mid P=N) = \frac{2}{5}$	$p(O=S \mid P=Y) = \frac{3}{9}$
1	1

$$P(P=Y \mid e) = \prod_i \frac{p(e_i \mid P=Y)\,p(P=Y)}{p(e_i)} \rightarrow p(e_i \mid P=y)p(P=y) + p(e_i \mid P=\neg y)p(P=\neg y)$$

MAP – assign the most probable category h_{MAP} given $e_1, \ldots, e_n$

$$h_{MAP} = \max_h P(h \mid e_1 \ldots e_n) = \max_n \frac{P(e_1 \ldots e_n \mid h)\,P(h)}{P(e_1 \ldots e_n)} \propto \max_n P(e_1 \ldots e_n \mid h)\,P(h)$$

Zero Frequency Problem

Add 1 to the numerator for ever attribute in frequency table so that probability can never be zero

Outlook	Play = No	Play = Ye		
Sunny	3 + 1	2 + 1 →	$p(O=S \mid P=N) = \frac{4}{8}$ (+1)	$p(O=S \mid P=Y) = \frac{5}{12}$ (+3)
Overcast	0 + 1	4 + 1 →	$p(O=O \mid P=N) = \frac{1}{8}$	$p(O=O \mid P=Y) = \frac{5}{12}$
Rainy	2 + 1	3 + 1	$P(O=R \mid P=N) = \frac{3}{8}$	$p(O=S \mid P=Y) = \frac{4}{12}$
			1	1

Missing Values

Missing val is numeric → Sample missing/new data x from $N(\mu_i, \sigma_i)$ for feature i

Classify the label for that new data $P(y|e_1=x, e_2, e_3 \dots e_n) \propto p(e_1=x, e_2 \dots e_n|y)\, p(y)$

Bayesian Network

Naive Bayes assumption of all feature being indep. to each other is too restrictive

Bayesian Network — conditional indep. among subset of feature

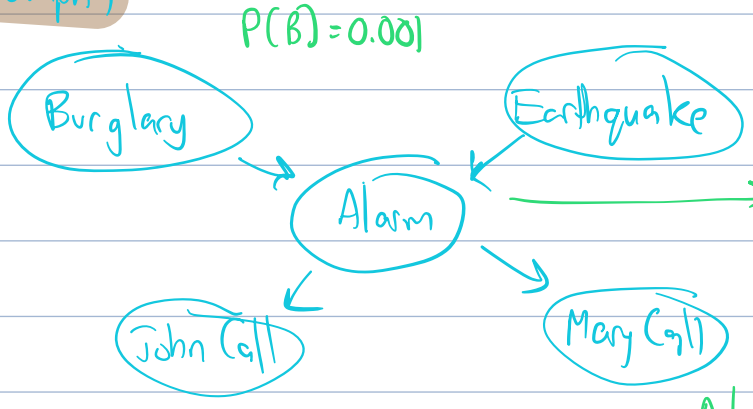
$$P(e_1, \dots, e_n) = \prod_{i=1}^{n} P(e_i | e_{i-1}, e_{i-2}, \dots e_1) = \prod_{i=1}^{n} P(e_i \mid \text{Parent}(e_i))$$

(Parent(e_i): dependent to e_i only)

$P(A,B,C,D,E) = P(A)\,P(B|A)\,P(C|A,B)\,P(D|A,B,C)\,P(E|A,B,C,D)$

$P(A|B,C)$ if C is conditionally indep of A given B, then $P(A|B,C) = P(A|B)$

DAG (Graph)



B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.9
F	0.05

A	P(M|A)
T	0.7
F	0.01

$P(J, M, A, \neg B, \neg E) = P(J|A)\,P(M|A)\,P(A|\neg B, \neg E)\,P(\neg B)\,P(\neg E)$

$= (0.9)(0.7)(0.001)(0.999)(0.998) = 0.000628$

Lecture 6: Decision Tree

Internal node – test on an attribute

Branch – subset of the tree from the outcome of the test

Leaf node – doesn't split any further, give class label or distribution

At each node, one attribute is chosen to split training example

Top-down Tree construction

- At start, all training example are at the root.
- Partition the example recursively by choosing one attribute each time

Bottom-up tree Pruning

- Remove subtree or branch from the bottom to top to improve the accuracy on new case (no overfitting)

Choosing Attribute to split

Information gain - choose attribute that result in greatest info gain

Info gain = info before split − info after split

gain(Outlook) = info([9,5]) − info([2,3], [4,0], [3,2])

Look at the gain for each attribute → choose attribute w/ the highest gain (purity)

$\text{Entropy}(p_1, \dots, p_n) = -\sum_i p_i \log_2(p_i)$ $p_i = p(y=c|t)$ => prob of picking class c in data set t

(use ln for the exam instead)

$\text{gain}(S, a) = \text{entropy}(S) - \sum_{v \in a} \frac{|S_v|}{|S|} \text{entropy}(S_v)$

($|S_v|$: # of value of specific type (Outlook = Sunny → 5); S_v: subset of data that have specific type of value and calculate entropy; $v \in a$: All possible type of value for feature a)

This is info gain after splitting tree with feature a

$i_E(t) = -p(y=\text{No}|\text{Outlook}=\text{Sunny}) \ln p(y=\text{No}|\text{O}=\text{Sunny}) - p(y=\text{Yes}|\text{O}=\text{Sunny}) \ln p(y=\text{Yes}|\text{O}=\text{Sunny}) = -\frac{3}{5}\ln\left(\frac{3}{5}\right) - \frac{2}{5}\ln\left(\frac{2}{5}\right)$

Splitting for ID Attribute → gain might be high (cus each group is pure) but then there's too many branches

Solution: Gain Ratio reduce bias on high branch attribute

- Take number and size of branch into account

Calculation intrinsic_info and $\text{gain Ratio}(S,a) = \frac{\text{gain}(S,a)}{\text{intrinsic\_info}(S,a)}$

$\text{info}[1,2,2] = -\frac{1}{5}\ln\left(\frac{1}{5}\right) - \frac{2}{5}\ln\left(\frac{2}{5}\right) - \frac{2}{5}\ln\left(\frac{2}{5}\right)$

info[1,2,2] = Entropy$\left(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right)$

so 3 branch, 1 data on 1 branch, 2 data on the other 2 branch

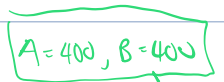
$\text{info}([2,3],[5,4]) = \frac{5}{14}\text{info}[2,3] + \frac{9}{14}\text{info}[5,4]$

$= \frac{5}{14}\left(-\frac{2}{5}\ln\left(\frac{2}{5}\right) - \frac{3}{5}\ln\left(\frac{3}{5}\right)\right) + \frac{9}{14}\left(-\frac{5}{9}\ln\left(\frac{5}{9}\right) - \frac{4}{9}\ln\left(\frac{4}{9}\right)\right)$

Gini Index

$Gini(p) = 1 - \sum_j p_j^2$ p_j = prob of picking class j

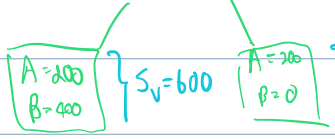
opt I)



opt I) $1 - \left(\frac{300}{400}\right)^2 - \left(\frac{100}{400}\right)^2 = 0.375$

opt II) left side $1 - \left(\frac{200}{600}\right)^2 - \left(\frac{400}{600}\right)^2 = 0.44\ldots$

opt II) right $\rightarrow 1 - \left(\frac{200}{200}\right)^2 - 0^2 = 0$



$S_V = 600$ $S_V = 200$

$\frac{|S_v|}{|S|}$ (same for both side)

$\frac{400}{800}(0.375) = 0.1875 \times 2 = 0.375$

$\frac{600}{800}(0.44) = 0.333\ldots$

$\frac{200}{800}(0) = 0$

total data on this split side

$\Delta i(t) = 0.5 - 0.375 \rightarrow 0.125$

$\Delta i(1) = 0.5 - 0.33 = 0.17 \rightarrow$ more gain!

DT w/ Numeric Attribute

Numerical Attr. can be use several time in DT while Nominal (class) Attri only once

Place split points halfway between the value eg. ..... 71 ; 72 ....

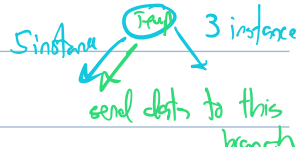
No : No

↳ split here < 71.5 and ≥ 71.5

$p(No | < 71.5)$ $p(No | \geq 71.5)$

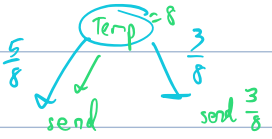
Handle missing value in training

1. Ignore data w/ missing value
2. Ignore attribute
3. Treat missing value as another nominal value (eg. label as "error")
4. Estimated missing value



Classifying data w/ missing value in the attri.

1. Follow the leader → send all the data w/ missing value on the attribute to the branch w/ most instance (training)
2. "Partition" the instance → Send down proportion of data to each branch based on the proportion of training instance



What can you do to make DT smaller and easier to interpret?

1. Pre Pruning (Early stop) 2. Post pruning 3. Limit depth of DT 4. use gain ratio (could help)

(reduce # of branch)

Lecture 7 : Data Preperation and Causal Inference

In DT : Preprunning and Post pruning

Preprunning — stop splitting the tree if $g(s,a) <$ threshold

Postprunning — Construct a complete DT first w/o threshold, then prune it back

— When prune → replace subtree w/ a single leaf node (remove all children of node but not the node itself)

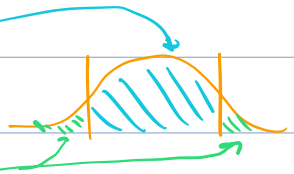
Prune the node if validation (hold out) error rate is reduce after pruning err(before prune) > err (after prune)

The observed error rate $f = \frac{\text{errors}}{n}$ follows Normal distribution

Central limit theorem - the standardised avg of any population of i.i.d RV X_i w/ μ_x and $\sigma^2 \sim N(0,1)$

Confidence Interval of Normal Dist.

% confidence interval for RV X w/ $\mu=0$ $= P(-z \leq X \leq z) = x$



and $P(-z \leq X \leq z) = 1 - 2P(z \geq X)$

tail distri.

Data Preperation

Data cleaning

1. Join relational data file together
2. Deal w/ missing entry → ignore data w/ missing entry
   → Treat missing entry as seperate value
   → fill missing entry w/ mean or median value
3. Conversion

   Convert Ordinal data (ordered attri (grade, size)) to numeric perserving natural order → ordinal encoding

   Convert nominal data (unordered) to binary dummy variable with one-hot encoding → Dummy variable encoding

   Discretization (Binning) - reduce number of values for a continuous attribute

   - Convert Numerical attr. to discrete attr. → group them into bucket
4. Build balance train&val sets (no imbalance set)
5. Subset Selection — Select subset of attribute that is the most relevant
   - Backward elimination - Use everything then remove attr. one by one
   - Forward selection - Choose one attr. at a time

Causal Inference

2 attr. are related but did one attr. have the effect on the other? (Correlation doesn't imply Causation)

Sample Selection bias → data collection process is bias

1) **Experimental studies** – Researcher intervene and observe what happen, indep. var are controlled

1.1) **Randomized Controlled Trials (RCT)** → (A/B Testing)

Randomized exp. where each subject is randomly assigned to a treated group or a control group in order to control for extraneous factors. Randomization mitigate the sample selection bias

Extraneous factor – variable that you're not investigating (not indep. var) that can potentially affect the result

1.2) **Quasi-Experiments**

Compare groups and measure effects w/o randomization of the subjects.

Assignment of subjects is not random → Selection bias is an issue.

2) **Observational Studies** – Researcher studies what occurs but doesn't change the subject

- Indep var is not under control of the research (opposite of Experimental study)
- Selection bias is a concern

2.1) **Cross-Selectional Study**

- Involve data collection from a population or subset of it at one specific point of time

eg. Sample of commuter on a given morning and study the mode of transportation

2.2) **Panel (or cohort) studies**

Observe a group of subjects over a span of time.

2.3) **Case-control study**

Study in which 2 existing groups differing in outcome are identified and compared on the basis of some supposed causal attribute

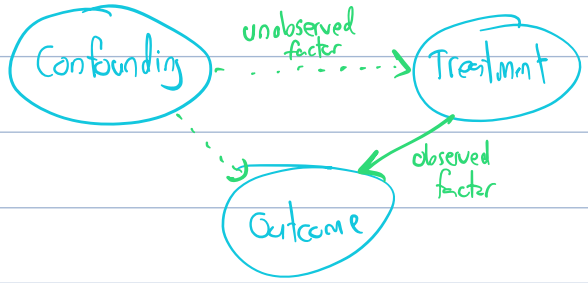
eg. Compare history of cancer patient vs. non-patient.

We want to identify causal, but there might be confounding variable

Confounding variable

is an extraneous var. that correlate (directly or inversely) w/ both dep and indep. var. in a way that "explain away" some or all of the correlation between the dep. and indep. var

- Confounding is part of exogeneity (one of assumption in Gauss-Markov)



How to deal w/ confounding var. (to identify causal effects)

1. RCT

Lab experiment → lead to low external validity (cannot explain data outside of observed population)

Create situation w/ desired conditions

Manipulate some var. while controlling others

Examine dependent variable

Field experiment → lead to high external validity

Research study in natural setting, manipulate some variable (no control), examine dep. var.

	Randomized Exp. (lead to high internal)	Quasi Exp. (lead to low internal) → more confounding var. from selection bias
Field	high internal / high external validity	low internal / high external validity
Lab	high internal / low external validity	low internal / low external validity

2. Fixed Effects Models → for panel study

Assume confounding var. has a fixed or constant relationship w/ the dep. variable across all observation

(So it can be subtract out w/ other data from diff group)

3. Difference in difference models (quasi-exp.)

We take a common trend assumption (treatment & control group has the same overall trend)

eg. Find causal effect on new law being introduce

City w/o new law (Control), Cities w/ new law (treatment)

Year	Treatment	Control
2019	9	10.2
2021	6.9	8.7

law take into effect during this time

We use control group to find the trend of the data

$$(Y_{t_2} - Y_{t_1}) - (Y_{c_2} - Y_{c_1}) = (6.9 - 9) - (8.7 - 10.2) = -0.6$$ → decrease after law

4. Propensity Score Matching (PSM) (Cross Selectional data)

Lecture 8 : Evaluation

Bias-Variance Trade off

Bias – Error from inaccurate prediction b/c it's too simple (not enough param)

Variance - error from sensitive to small fluctuation. Too many params.

Underfitting - model is too simple to represent all relevant characteristics

Over fitting – model is too complex and fit irrelevant noise/ characteristic

Resampling Method

- Holdout procedure
  - reserve data for test set and use remaining for training and validation
- Stratified holdout
  - guarantee that classes are proportionally represented in the test/train set
- Repeated holdout

  Randomly select test set (not val set like cross Val) and avg the error rate estimate

K-fold Cross-Validation

- fixed # of k partition of the data (fold)
- Each fold, select one partition as a validation set and the rest as training set (each partition is use for validation only once)
- Error rate is estimate by avg of error rate of all k fold

Use paired t-test to check if estimate error of model is "significantly" difference from esti. error rate of another model

Bootstrap

1. Draw a random sample of size n and calculate statistic (error rate, mean) with these data
2. Repeat it k many times, getting error = $(e_1, e_2, e_3, \ldots e_k)$ $\mu = (\mu_1, \mu_2, \mu_3, \ldots \mu_k)$
3. Calculate variance of your statistic to learn about population statistics
4. Calculate confidence interval of your original data

Measuring Error

		Predicted T	Predicted F
Actual	T	TP	FN
	F	FP	TN

$$\text{Error rate} = \frac{\#\text{ of error}}{\#\text{ of instance}} = \frac{(FN+FP)}{N}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{False Alarm} = \frac{FP}{FP+TN}$$

Gain Curve

% of population vs # of target (that you're interested in)

$$\text{gain curve} = \frac{\#\text{ of target for \% of population}}{\text{Total \# of target}} = \frac{n}{N}$$

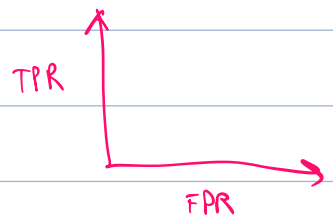
Lift Curve

$$\text{Lift}(q) = \frac{\text{Gain}(q)}{q}$$

q = % of sample size

ROC Curve

TP rate (Recall) vs False Positive



(look at exercise on how to calculate TPR and FPR)

With gain curve & lift curve, you can see which model perform best at low sample size (like it predict 80% correct w/ 50% data vs predict 80% at 80% data)

Lecture 9 : Ensemble Methods

Idea - Combining Multiple Models

Advantage

- Often improve Predictive Perf.

Disadvantage

- Output is hard to analyze

Bagging

- Aim to reduce variance
- Each Model receive equal weight
- How it work :
  1) Sample several (new) training set of size n from the (original) training data (Bootstrap)
  2) Built a classifier (model) for each training set
  3) Combine classifier's prediction
- Advantage
  - Can be apply to numeric prediction & classification
  - Help perf if data is noisy
  - The more model, the better perf (w/ diminishing return)
  - Can be parallelized cus model can be create independently

Random Forest → Combine Bagging with Decision Tree

→ Randomized training data (Bootstrap) and feature!

(randomize the feature selection for each DT for bagging)

For each data x_i, return class that was predicted most often for all DT.

Each tree has high variance & low bias → Agg all the tree help reduce variance

Boosting

Idea: Combine several weak learners into a strong learner

New models are influenced by perf. of previously built model ("new" model try to fix the problem of previous model)

- Increase capability of the model, minimizing the bias (training error)
- Weights models according to performance

eg. Ada Boost, XG Boost

Bias-Variance Tradeoff!

Bagging reduces variance of low-bias models

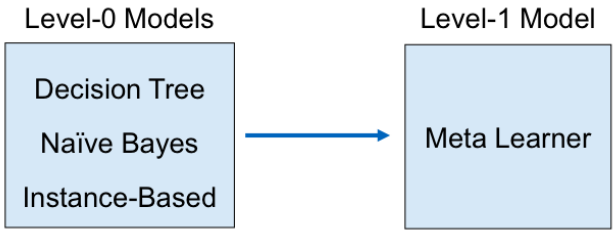
- Low-bias models are "complex" and unstable
- Bagging averages them together to create stability

Boosting reduces bias of low-variance models

- Low-variance models are simple with high bias
- Boosting trains sequence of models on residual error → sum of simple models is accurate

Stacking

Prediction from each homogenous classifiers are used as an input into a meta learner which attempt to combine prediction to create a final best prediction



Clustering – Unsupervised learning

- Item in same cluster are similar (high similarity)
- Item from diff cluster are different (low similarity)

1. Hierarchical Clustering

Bottom Up

Start from each data has its each cluster

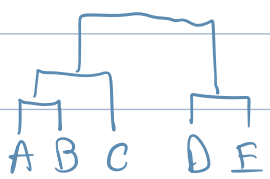
At each step, merge 2 closest clusters

Top down

Start w/ all data in a cluster

Seperate into 2 cluster at a time

use Dendogram



Use Dissimilarity Matrices (calculate based on distance) to find closest cluster

MST – Minimum Spanning Tree

- Cheapest path from data A to B → A and B is in same cluster
- Each cheapest path between 2 data/cluster is a join in the cluster.
- Create a dendogram from the cluster

K-Means Clustering

- work w/ numeric only

How it work:

1) initialise k cluster centroid in random
2) Assign every data to the nearest cluster (Euclidean - L_2, Manhattan - L_1, etc.)
3) Move the centroid to the mean of its assigned data
4) Repeat 2 & 3.

Result can vary significantly depending on initialised centroid

Probability-Based Clustering (Gaussian Mixture Model)

Mixture Model = Model each cluster w/ prob. distribution (Gaussian)

Each probability distribution give prob of the data belonging to each cluster $p(x_i) = (c_1, c_2, \ldots, c_k)$

Expectation Maximization (algorithm to optimize GMM)

Start w/ initial param $\sigma_A, \sigma_B, \mu_A, \mu_B, P(A)$

Iterate between E and M

Expectation - calculate probability of data being on each cluster $\gamma_t(z_{ik}) = p(z_{ik} | x_i, \pi, \mu_k, \sigma_k)$ (prob of cluster k given x_i)

Maximization - Find the maximum likelihood of each param from $\gamma_t(z_{ik})$ - $\pi_k^*, \mu_k^*, \sigma_k^*$

Use the recalculate param to calculate expectation again $\gamma_{t+1}(z_{ik})$

Use recalculate expectation $\gamma_{t+1}(z_{ik})$ to find MLE param again.

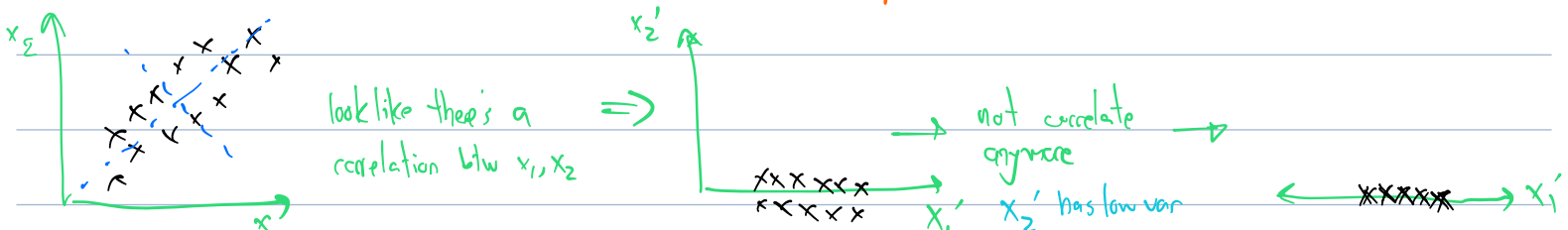
Lecture 10: Dimensionality Reduction

PCA - Convert a set of possibly correlated variables into a set of values of linearly uncorrelated var.

This can help combat multicollinearity in the regression analysis

Transform the coordinate system to the principle component coordinate system

eigenvector $\vec{v}$ are direction that doesn't rotate when transform, only scaling by λ



, can get rid

How to transform data w/ PCA

Step 1: Compute zero mean data (compute mean for each column), then $\tilde{X} = X - \bar{X}$

Step 2: Calculate Covariance Matrix of $\tilde{X}$

$\Sigma_X = \frac{X^T X}{N-1}$ or $\Sigma_X = \begin{pmatrix} Var(x_1) & Cov(x_1,x_2) \\ Cov(x_2,x_1) & Var(x_2) \end{pmatrix}$

$Var(x_i) = \frac{\sum_j (x_{ji} - \bar{x}_i)^2}{N-1}$ → if X is zero mean → $Var(x_i) = \frac{\sum_j x_{ji}^2}{N-1}$

$Cov(x_i, x_j) = \frac{\sum_k (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{N-1} = \frac{\sum_k x_{ki} x_{kj}}{N-1}$

Step 3: Calculate v and λ of $\Sigma_X \rightarrow \Sigma = \Phi \Lambda \Phi^T$

Final data: $Z = X \cdot \Phi$ where Φ = eigenvector matrix

Can leave out dimension w/ small variance → have enough $\vec{v}$ to cover 80-90% variance

Reconstruct data $X = Z \cdot \Phi^T$ + original mean

SVD

$A = U \Sigma V^T$ where $\Sigma(A) = \sqrt{\lambda(A^T A)}$

V is the principle component coordinate system

$U\Sigma$ = coordinate of A in the new coordinate system · scaling

$Z = AV = U\Sigma$

PCA is not feature selection! it's feature extraction.

Feature selection, you reduce # of dim by throwing away irrelevant information (use same data but throw away some feature)

Feature Extraction (PCA), reduce # of dim by transforming data (projecting) to low dim data, thus still maintaining the same information

Lecture 11: Convex Optimization & NN

$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$ $x \in \mathbb{R}^n$ = gradient of f = direction of steepest ascent

$-\nabla f(x)$ = steepest descent

Convex

f is convex iff $\forall \lambda \in [0,1] \quad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$

$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in \text{dom}(f)$, $\text{dom}(f)$ is convex

f is convex iff $\nabla^2 f(x) \geq 0$ or $x^T H(x, y, \ldots) x \geq 0$ (Hessian is PSD)

f is convex $\iff$ $H_f(x)$ is PSD

if $\nabla f(x) = 0$, $H_f(x)$ is Positive Definite $\rightarrow$ x isolate local min & λ of $x > 0$

$\nabla f(x) = 0$, $H_f(x)$ is Negative Definite $\rightarrow$ x isolated local max & λ of $x < 0$

$\nabla f(x) = 0$, $H_f(x)$ is indefinite $\rightarrow$ λ of x has both < 0 and > 0 $\rightarrow$ saddle point

$\nabla f(x) = 0$, $H_f(x)$ is PSD $\rightarrow$ inconclusive λ of $x \geq 0$

but

x is local min $\rightarrow$ $H_f(x)$ is PSD

x is local max $\rightarrow$ $H_f(x)$ is ~~PSD~~ NSD

Gradient Descent

$x^{(k+1)} = x^{(k)} - \alpha \nabla f(x^{(k)})$ change over iteration

α can be a line search (expensive) or fixed α or α_t

Start at $x^{(1)}$ $\rightarrow$ calculate $\nabla f(x^{(1)})$ using $x^{(1)}$ บันทึกไว้ก่อน

then $x^{(2)} = x^{(1)} - \alpha \nabla f(x^{(1)})$

if α not given, then $\alpha^* = \arg\min_{\alpha} f(x^{(1)} - \alpha \nabla f(x^{(1)})) = \frac{\partial}{\partial \alpha} f(x^{(1)} - \alpha \nabla f(x^{(1)})) =$ input for func.

Momentum Gradient Descent

Use momentum $x^{(k+1)} = x^{(k)} - \alpha V_t$ and $V_t = \beta \nabla f(x^{(k)}) + (1-\beta) V_{t-1}$

Increase the step if the direction go in the same direction (accumulate gradient)

Multiple Linear Regression Model

+1 for bias

Estimated equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_d x_d = X\hat{\beta} + e \qquad Y \in \mathbb{R}^{n\times 1} \quad X \in \mathbb{R}^{n\times(d+1)} \quad \hat{\beta} \in \mathbb{R}^{(d+1)\times 1} \quad \varepsilon \in \mathbb{R}^{n\times 1}$$

(e: error)

Loss function of MLR = RSS

$$RSS = e^Te = (y - X\hat{\beta})^T(y - X\hat{\beta}) \rightarrow \text{convex b/c let } V = y - X\beta^T \quad x^TV^TVx - (Vx)^T(Vx) = \|Vx\|_2^2 \geq 0$$

Ordinary Least square estimation → find param that will minimize the loss

$$\beta_{LS} = (X^TX)^{-1}X^Ty \qquad \text{so} \qquad \hat{y} = X(X^TX)^{-1}X^Ty$$

Do OLS Estimation via GD

Use mean square error func $= \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$

$$\frac{\partial L}{\partial \beta_0} = -\frac{2}{n}\sum_i (y - y_i) \qquad \frac{\partial L}{\partial \beta_1} = -\frac{2}{n}\sum_i x_i(y - \hat{y}_i)$$

$$\beta_0 = \beta_0 - \alpha\frac{\partial L}{\partial \beta_0} \qquad \beta_1 = \beta_1 - \alpha\frac{\partial L}{\partial \beta_1}$$

Best Linear unbiased estimator

Recap Gauss-Markov Theorem : OLS estimator give the BLUE of the coefficient

Unbiased $= E(\hat{\beta}_j) = \beta_j$ Best = give lowest variance compare to other linear unbiased estimator

But often, there exist biased estimator that give lower error/variance

Regularization – estimator will be more biased but lower var

Ridge Regression (L_2)

$$L(\beta) = \sum_i (y_i - \hat{y}_i) + \lambda \sum_j \beta_j^2$$

β will shrink all equally

Lasso (L_1)

$$L(\beta) = \sum_i (y - \hat{y}_i) + \lambda \sum_j |\beta_j|$$

β will be more sparse (more 0's entry) b/c less important feature get set to 0, Lasso perform a feature selection. Also, make model

simpler to interpret.

Logistic Regression

Likelihood $L = \prod_i p^{y_i}(1-p)^{1-y_i}$ let $p(y|x) = \sigma(\beta_0+\beta_1 x) = \sigma(z)$

Log Likelihood $LL = \sum_i y_i \ln p_i + (1-y_i)\ln(1-p_i)$

Negative LL $= \sum_i -y_i \ln p_i - (1-y_i)\ln(1-p_i)$

Is NLL convex?

$\beta_0+\beta_1 x$ is affine func → convex

$\sigma(\beta_0+\beta_1 x)$ is convex

is $f(g(z)) = -\ln \sigma(\beta_0+\beta_1 x)$ convex? ($\ln(1+\exp(-(\beta_0+\beta_1 x))$)

★ if $f''(g(z))$ is ≥ 0 (f is convex) and $f'(g(z))$ is monotonically increasing and g is convex ($g''(z) \geq 0$)

then $f(g(z))$ is convex

∴ You can find the minimum of NLL

Prove

$f'(g(z)) \cdot g'(z) = f'(g(z)) \cdot g''(z) + g'(z) f''(g(z)) \cdot g'(z)$

$= f'(g(z)) \cdot g''(z) + g'^2(z) f''(g(z)) \geq 0$

Lecture 12 : NN

Risk function R is a expectation of a loss function (avg loss per data point)

Loss func – loss of the whole dataset → $L(\theta) = -\sum_{i=1}^{n} p_i \log(\hat{y}_i) + (1-p_i)\log(1-\hat{y}_i)$

$R(\theta) = \frac{1}{N}\sum_n L(y_n, g(\theta, x_n))$ if Loss func = MSE, → $L(\theta) = -\frac{1}{n}\sum_{i=1}^{n} p_i \log(\hat{y}_i) + (1-p_i)\log(1-\hat{y}_i)$ if BCE

then $R(\theta) = \frac{1}{N}\sum_n \left(\frac{1}{2}(y_n - g(\theta^T x_n))^2\right)$

★ We use Risk func (avg w/ $\frac{1}{n}$) here in this class → In ML we use total loss

GD

Update w/ small α → very small step, might get stuck in saddle point

w/ large α → oscillate

A 1 Layer NN is similar to Logistic Regression. But NN can stack many layer = more complex

1 Layer NN: no hidden layer, output layer = 1 layer

Input layer = # of feature of data

Output layer = # of class

let model consist of H_1 = hidden node for layer 1, H_2 = hidden node for layer 2, Y = output node, X = input node

# of weight in node : $(X+1)\times(H_1) + (H_1+1)\times(H_2) + (H_2+1)\times(Y)$

(+1 = bias)

Backpropagation

Calculate gradient of params of NN w/ respect to loss func

$$w_{ij}^{(t+1)} = w_{ij}^{t} - \alpha \underbrace{\frac{\partial R}{\partial w_{ij}}}_{\frac{1}{n}\sum_i L(y, g(x,w))}$$

$$w_{jk}^{(t+1)} = w_{tk}^{t} - \alpha \frac{\partial R}{\partial w_{tk}}$$

$$W^{(t+1)} = W^{(t)} - \alpha \nabla_W L(W^{(t)})$$

Matrix Notation for 2 layer NN (1 hidden layer and σ activation func)

let $Z^{(1)} = XW^{(1)} + b^{(1)}$

$A = \sigma(Z^{(1)})$

$Z^{(2)} = AW^{(2)} + b^{(2)}$

$\hat{Y} = \sigma(Z^{(2)})$

$$\hat{Y} = \sigma(\sigma(XW^{(1)} + b^{(1)})W^{(2)} + b^{(2)})$$

Mini Batch

with n data, create m batch, each w/ $\frac{n}{m}$ data. For each batch, do forward pass, calculate loss, backward pass and update param. Update param m time per epoch.

Stochastic Gradient

Minibatch w/ only one data for each batch, calculate loss of 1 sample and do backprop and update the parameter. So if n=500, do 500 update per epoch.

Both provide noisy gradient which can be good for NN.

Hidden layer do feature extraction (transform info, didn't get rid of info (feature selection)

Universal Approximation Theorem

NN w/ one hidden layer and non linear activation func can approx. any func.

Lecture 13: Reinforcement Learning

What is RL?

RL incorporate time into learning.

It learn to make sequential decision in an environment which maximize "rewards"

- Reward tell how well our agent behaved (reward can be positive or negative)

How RL diff from Supervised and Unsupervised L?

- No supervision
- Feedback can be delayed
- Time matter (sequential data)
- Not i.i.d. (Action affect subsequent data)

Markov Process - Observe sequene of states or a chain

System change between states according to some dynamics

Markov Process iff future system dynamic from any state depend only on current state.

"Future in independent of the past, given the present)

Markov Decision Process - Markov Process w/ action and reward

(if only 1 action exist for each state and all reward is same, than it becomes Markov Process)