TA Review for Exam 1

Jared and Sai

Today: Guided Practice

1. TA’s: Present a problem
2. You: Try to solve it on your own
3. TA’s: Review the needed material
4. You: Try the problem a second time with the reviewed tools
5. Together: Find the solution

Decision Trees - a few notes

Chronological: time proceeds from left to right

Probabilities are conditional on everything that's already observed (in other words, conditional upon everything to its left)

Perfect information implies 0's and 1's for probabilities

Decision Trees

You’re the founder of a hot tech startup company, and you’re ready to sell your shares. A private equity firm gives you an offer of $2M. Alternatively, you could go public, which has a 10% chance of yielding $5M and a 90% chance at $1M. You have a friend from McCombs who specializes in IPO consulting. She guesses 75% of successes correctly, and 85% of failures correctly. Based off her advice, you could then choose your sale strategy, but her advice costs $0.1M. What should you do?

Decision Trees

You're the founder of a hot tech startup company, and you're ready to sell your shares. A private equity firm gives you an offer of $2M. Alternatively, you could go public, which has a 10% chance of yielding $5M and a 90% chance at $1M. You have a friend from McCombs who specializes in IPO consulting. She guesses 75% of successes correctly, and 85% of failures correctly. Based off her advice, you could then choose your sale strategy, but her advice costs $0.1M. What should you do?

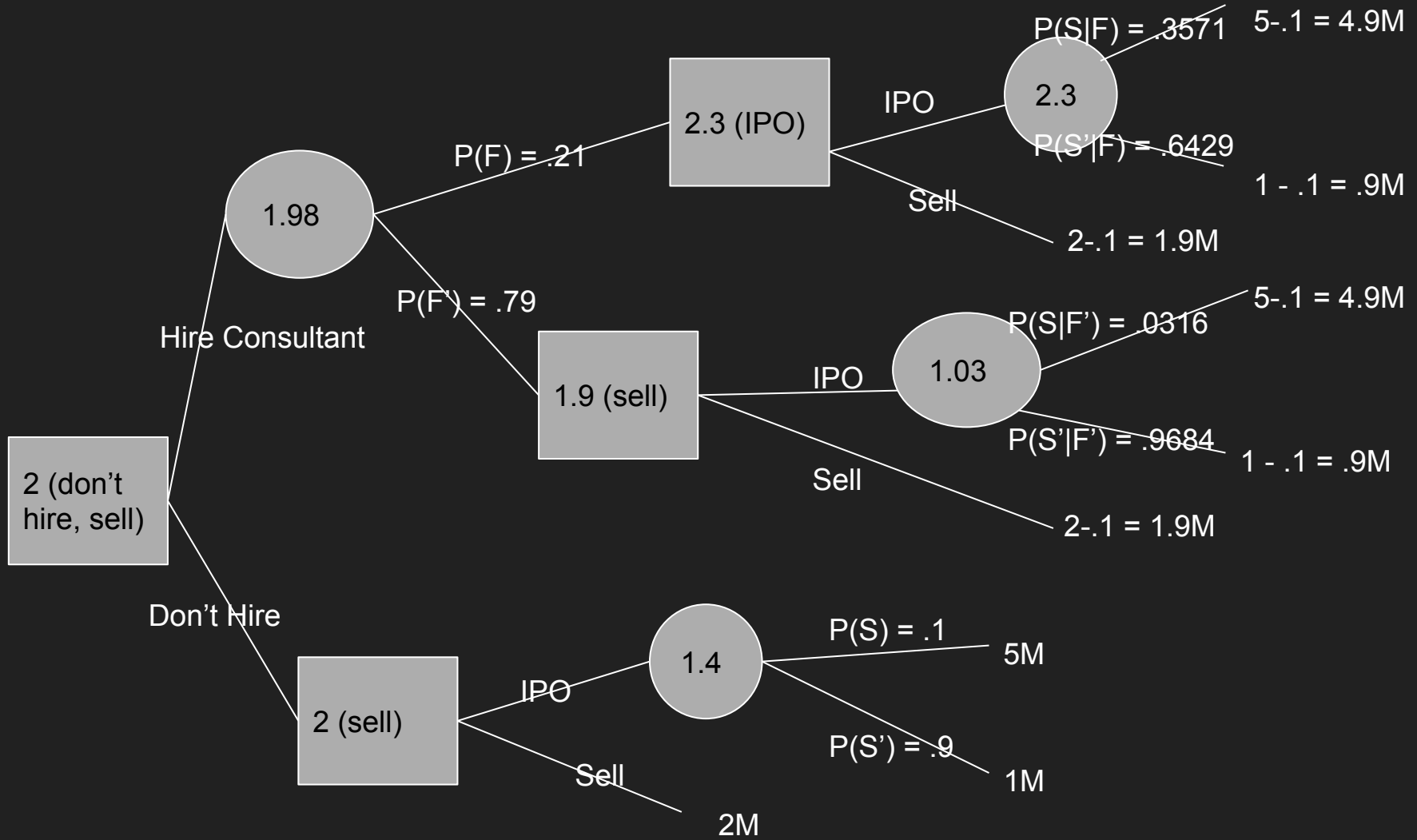
Remember it's chronological: first you must choose to hire the consultant or not, then where to sell.

Decision Trees

Solution: let F=forecasted success, and S=actual success ($5M), and F'=forecasted failure and S'=actual failure ($1M).

We're given P(S) = 0.1, P(F|S) = 0.75, P(not F| not S) = 0.85.
Then
P(F) = P(F|S)P(S) + P(F|not S)P(not S) = .75*.1 + .15*.9 = .21
P(not F) = 1 - P(F) = .79
P(S|F) = P(F|S)P(S)/P(F) = .75*.1/.21 =.3571
P(not S|F) = 1 - P(S|F) = 1 - .3571 = .6429
P(S|not F) = P(not F|S)P(S)/P(not F) = .25*.1/.79 = .0316
P(not S|not F) = 1 - P(S|not F) = 1 - .0316 = .9684

Now we have all the needed probabilities for our tree


2 (don't hire, sell)
Hire Consultant
Don't Hire
1.98
P(F) = .21
P(F') = .79
2.3 (IPO)
IPO
Sell
2.3
P(S|F) = .3571
P(S'|F) = .6429
5-.1 = 4.9M
1 - .1 = .9M
2-.1 = 1.9M
1.9 (sell)
IPO
Sell
1.03
P(S|F') = .0316
P(S'|F') = .9684
5-.1 = 4.9M
1 - .1 = .9M
2-.1 = 1.9M
2 (sell)
IPO
Sell
1.4
P(S) = .1
P(S') = .9
5M
1M
2M

Probability #1

Suppose (for simplicity) that the price of IBM's stock can increase $1 or decrease $1. Likewise for General Motors' stock. Let X be the change in the price of IBM's stock and let Y be the change in the price of General Motors' stock. Suppose that (X, Y) have the following joint distribution.

		Y	
		-1	+1
X	-1	0.25	0.15
	+1	0.05	0.55

a. Marginal probabilities for X? Marginal probabilities for Y?
b. What is the probability that the stock price of IBM goes up by $1?
c. If you know that GM's stock price will fall today, what is the probability that the stock price of IBM goes up by $1?

Probability #1 - Principles

Recall that two-way tables contain the joint probabilities. The sums of the rows and sums of the columns are the marginal probabilities (in the margins!). Also, note the definition of conditional probability: P(A|B) = P(A and B)/P(B).

		Y (GM)	
		-1	+1
X (IBM)	-1	0.25	0.15
	+1	0.05	0.55

	B	Not B	
A	P(A and B)	P(A and not B)	P(A)
Not A	P(not A and B)	P(not A and not B)	P(not A)
	P(B)	P(not B)	1.0

a. Marginal probabilities for X? Marginal probabilities for Y?
b. What is the probability that the stock price of IBM goes up by $1?
c. If you know that GM's stock price will fall today, what is the probability that the stock price of IBM goes up by $1?

Probability #1 - Solution

a. Marginal probabilities for X? Marginal probabilities for Y?
   i. Solution: sum the rows and columns
b. What is the probability that the stock price of IBM goes up by $1?
   i. P(IBM+1) = 0.6, as shown in table
c. If you know that GM's stock price will fall today, what is the probability that the stock price of IBM goes up by $1?
   i. Use the definition of conditional probability

		Y (GM)		
		-1	+1	
X (IBM)	-1	0.25	0.15	0.4
	+1	0.05	0.55	0.6
		0.3	0.7	1.0

$$P(IBM+1 \mid GM-1) = P(IBM+1 \text{ and } GM-1)/P(GM-1)$$
$$= 0.05/0.3$$
$$= 0.1667$$

Probability #2

a. Are changes in IBM stock price independent of changes in GM stock price?
b. What is the probability that either stock price increases?

		Y (GM)		
		-1	+1	
X (IBM)	-1	0.25	0.15	0.4
	+1	0.05	0.55	0.6
		0.3	0.7	1.0

Probability #2 - Principles

Recall that independence implies that

P(A and B) = P(A)P(B)

		Y (GM)		
		-1	+1	
X (IBM)	-1	0.25	0.15	0.4
	+1	0.05	0.55	0.6
		0.3	0.7	1.0

P(A or B) = P(A) + P(B) - P(A and B)

Also:

P(A or B) = P(A and B) + P(A and not B) + P(not A and B)

Probability #2 - Solution

a. Are changes in IBM stock price independent of changes in GM stock price?
   i. P(IBM+1 and GM+1) = 0.55
   ii. P(IBM+1) = 0.6, P(GM+1) = 0.7
   iii. 0.6*0.7 = .42, not 0.55, so cannot be independent

b. What is the probability that either stock price increases?
   i. P(IBM+1 or GM+1)
   = 0.05 + 0.55 + 0.15 = 0.75
   ii. Or:
   0.7 + 0.6 - 0.55 = 0.75

		Y (GM)		
		-1	+1	
X (IBM)	-1	0.25	0.15	0.4
	+1	0.05	0.55	0.6
		0.3	0.7	1.0

Probability Rules Summary

Marginal: P(A)

Joint: P(A and not B)

Conditional:
P(not B|A) = P(A and not B)/P(A)
Note that this means:
P(A and not B) = P(not B|A)P(A)

Independence implies
P(A)P(B) = P(A and B)

P(A or B) = P(A) + P(B) - P(A and B)

P(A) = P(A and B) + P(A and not B)
P(A) = P(A|B)P(B)+ P(A|not B)P(not B)

Bayes Rule:
P(B|A) = P(A|B)P(B)/P(A)

Bayes Rule (expanded denominator):

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}.$$

Probability #3

A cancer screening detects cancer in 95% of cancer patients, but it also detects cancer in 10% of cancer-free patients. It is estimated that 0.5 % of the population suffer from cancer. Suppose a random individual is screened. Calculate the following probabilities:

a. that the test result will be positive (detect cancer)
b. that, given a positive result, the person is a sufferer
c. that, given a negative result, the person is a non-sufferer
d. that the person will be misclassified

Probability #3 - Solution

A cancer screening detects cancer in 95% of cancer patients, but it also detects cancer in 10% of cancer-free patients. It is estimated that 0.5 % of the population suffer from cancer. Suppose a random individual is screened. Calculate the following probabilities:

T ~ Test Positive, S ~ Sufferer, M ~ Misclassified, then, P(T|S)= 0.95, P(T|S')=0.10

P(S) = 0.005, P(T|S)= 0.95, P(T|S')=0.10

a. P(T) = P(T|S)*P(S) + P(T|S')*P(S') = 0.95*0.005 + 0.1*0.995 = 0.10425
b. P(S|T) = P(T|S)*P(S)/P(T) = 0.95*0.005/0.10425 = 0.0455
c. P(S'|T') = P(T'|S')*P(S')/P(T') = 0.9*0.995/(1-0.10425) = 0.9997
d. P(M) = P(T & S') + P(T' & S) = P(T|S')*P(S') + P(T'|S)*P(S) = 0.09975

Probability #3 - Solution

T ~ Test Positive, S ~ Sufferer, M ~ Misclassified, then, P(T|S)= 0.95, P(T|S')=0.10

P(S) = 0.005, P(T|S)= 0.95, P(T|S')=0.10

Alternatively, first fill out the two-way table, then solve:
P(T and S) = P(T|S)P(S) = 0.95*.005 = .00475
P(T and S') = P(T|S')P(S') = .1*.995 = .0995
Then fill in so rows/columns sum appropriately

	S	S'	
T	.00475	.0995	.10425
T'	.00025	.8955	.89575
	.005	.995	

a. P(T) = 0.10425
b. P(S|T) = P(S and T)/P(T) = .00475/.10425 = .04556
c. P(S'|T') = P(S' and T')/P(T') = .8955/.89575 = 0.9997
d. P(M) = P(T & S') + P(T' & S) = .00025 + .0995 = 0.09975

Distributions #1 - Discrete Random Variable

Let’s play a game. You flip two coins. If both are heads, you get $2. If both are tails, you lose $5. If mixed heads/tails, then you get $1.

a. How much do you expect to win if you play the game once?
b. What is the standard deviation of winnings?

Distributions #1 - Principles

Let's play a game. You flip two coins. If both are heads, you get $2. If both are tails, you lose $5. If mixed heads/tails, then you get $1.

a. How much do you expect to win if you play the game once?
b. What is the standard deviation of winnings?

For discrete variables, $E(X) = \Sigma x P(x)$ and $Var(X) = \Sigma P(x)[x-E(x)]^2$

Also, standard deviation is the square root of the variance

Distributions #1 - Solution

Let's play a game. You flip two coins. If both are heads, you get $2. If both are tails, you lose $5. If mixed heads/tails, then you get $1.

a. How much do you expect to win if you play the game once?
   i. $E(X) = 0.25*(2) + 0.25*(-5) + 0.5*(1) = -0.25$
b. What is the standard deviation of winnings?
   i. $Var(X) = 0.25*(2-^{-}0.25)^2 + 0.25*(-5-^{-}0.25)^2 + 0.5*(1-^{-}0.25)^2 = 7.6875$
   ii. $SD(X) = sqrt(Var(X)) = 2.7726$

Distributions #2 - Normal Random Variable

Annual car sales at a local dealership roughly follow a normal distribution. They average 500 sales per year, and most years (about 95%) they sell between 400 and 600 cars.

a. What is the standard deviation of their annual sales?
b. What is the probability that on a randomly selected year they sell between 456 and 654 cars?

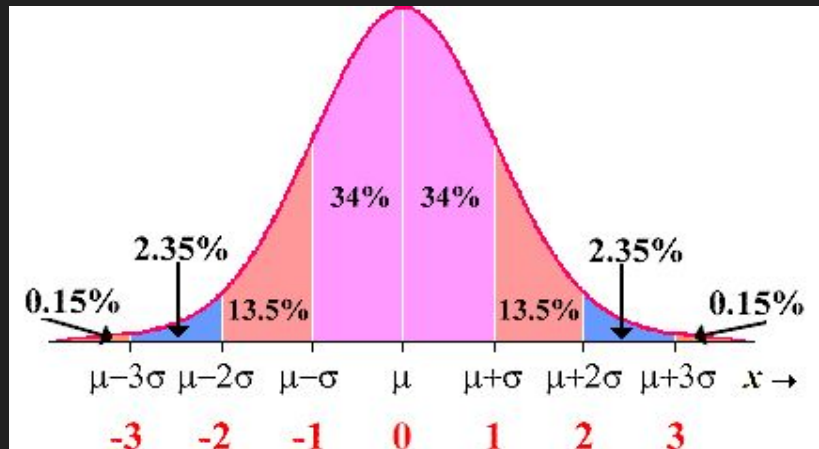
Distributions #2 - Principles

Annual car sales at a local dealership roughly follow a normal distribution. They average 500 sales per year, and most years (about 95%) they sell between 400 and 600 cars.

a. What is the standard deviation of their annual sales?
b. What is the probability that on a randomly selected year they sell between 456 and 654 cars?

Recall the 68-95-99.7 rule (approximation!)

Also recall that pnorm("value",mu,sigma) gives P(X<value) for a Normal(mu,sigma^2) distribution



Distributions #2 - Normal Random Variable

Annual car sales at a local dealership roughly follow a normal distribution. They average $500,000 in sales per year, and most years (about 95%) they sell between $400,000 and $600,000.

a. What is the standard deviation of their annual sales, in thousands?
   i. According to the 68-95-99.7 rule, about 95% of the distribution is within +/- 2 standard deviations of the mean. (600-400)/4 = $50
b. What is the probability that on a randomly selected year they sell between $456,000 and $654,000?
   i. P(456 < X< 654) = P(X<654) - P(X<456) =
   ii. = pnorm(654,500,50) - pnorm(456,500,50) = 0.8095

Regression #1

homes = read.csv('https://raw.githubusercontent.com/brianlukoff/sta371g/master/data/houses.csv')

This dataset contains randomly selected real estate transactions for houses sold. Create a regression model for predicting the sale price from the Living area.

a. Is living area a significant predictor of sale price?
b. How much of the variation in sale price is predicted by living area?
c. How much do you expect sales price to increase for 100 extra square feet of living area?
d. What is the forecasted sale price of a 2,000 sq. ft. house?

Regression #1 - Principles

```
Call:
lm(formula = homes$Price ~ homes$Living.Area)

Residuals:
    Min      1Q  Median      3Q     Max
-277022  -39371   -7726   28350  553325

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      13439.394   4992.353   2.692  0.00717 **
homes$Living.Area  113.123      2.682  42.173  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69100 on 1726 degrees of freedom
Multiple R-squared:  0.5075,    Adjusted R-squared:  0.5072
F-statistic:  1779 on 1 and 1726 DF,  p-value: < 2.2e-16
```

Regression #1 - Solution

Create a regression model for predicting the sale price from the Living area.

a. Is living area a statistically significant predictor of sale price?
   i. The p-value is < 2e-16, so yes
b. How much of the variation in sale price is predicted by living area?
   i. This is asking for R^2, which is 0.5075, so 50.75%
c. How much do you expect sales price to increase for 100 extra square feet of living area?
   i. The slope on living area is 113.1225, so for 100 extra sqft we'd expect a $11,312.25 increase
d. What is the forecasted sale price of a 2,000 sq. ft. house?
   i. The intercept is 13439.3940, so 13439.3940+113.1125*2000 = $239,684.48

Regression (#2) Assumptions

Does this model (predicting the sale price from the living area) satisfy the assumptions for regression models?

Regression (#2) Assumptions - Principles

Does this model (predicting the sale price from the living area) satisfy the assumptions for regression models?

L - Linear

I - Independent

N - Normal

E - Equal variance (homoscedasticity)

Regression (#2) Solution

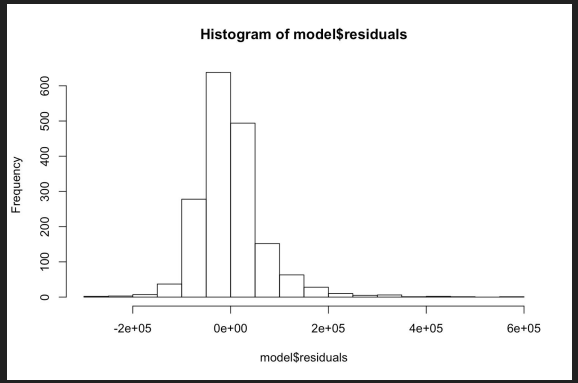
Does this model (predicting the sale price from assumptions for regression models?

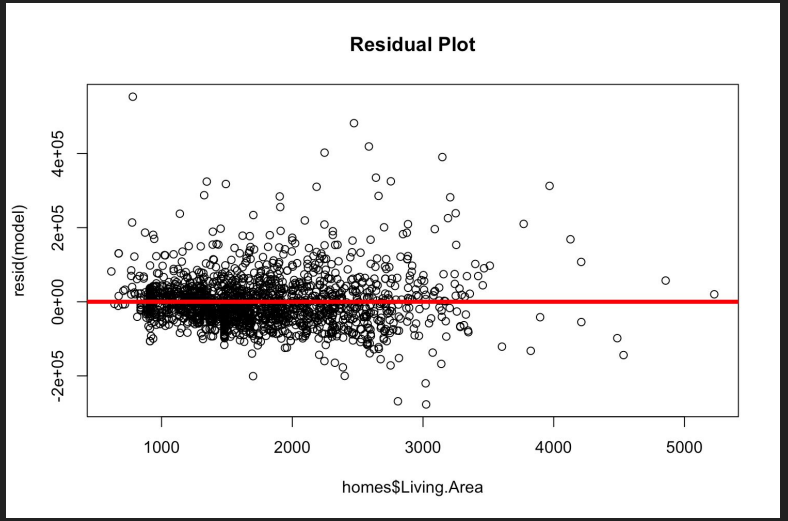
L - Linear

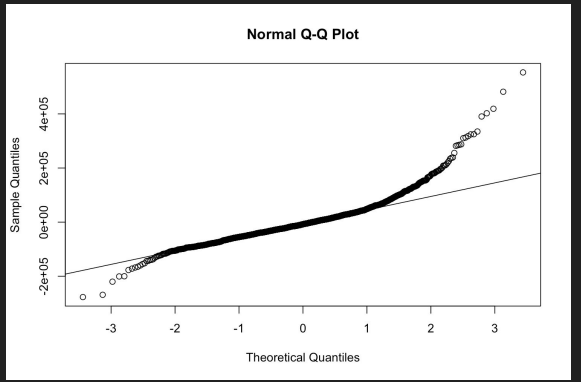
I - Independent

N - Normal

E - Equal variance (homoscedasticity)







R basics

Save your work in R!

Quantitative vs Categorical data

Subsetting data

Graphics: histogram, scatterplot

Other material

Population vs sample

Recall hypothesis testing

t.test() function: t-tests and confidence intervals

pnorm, pt

p-value: P(seeing data this extreme in our sample | Ho is true)