Lecture 19 – Review, Conclusion

DSC 80, Winter 2024

Announcements

hi!

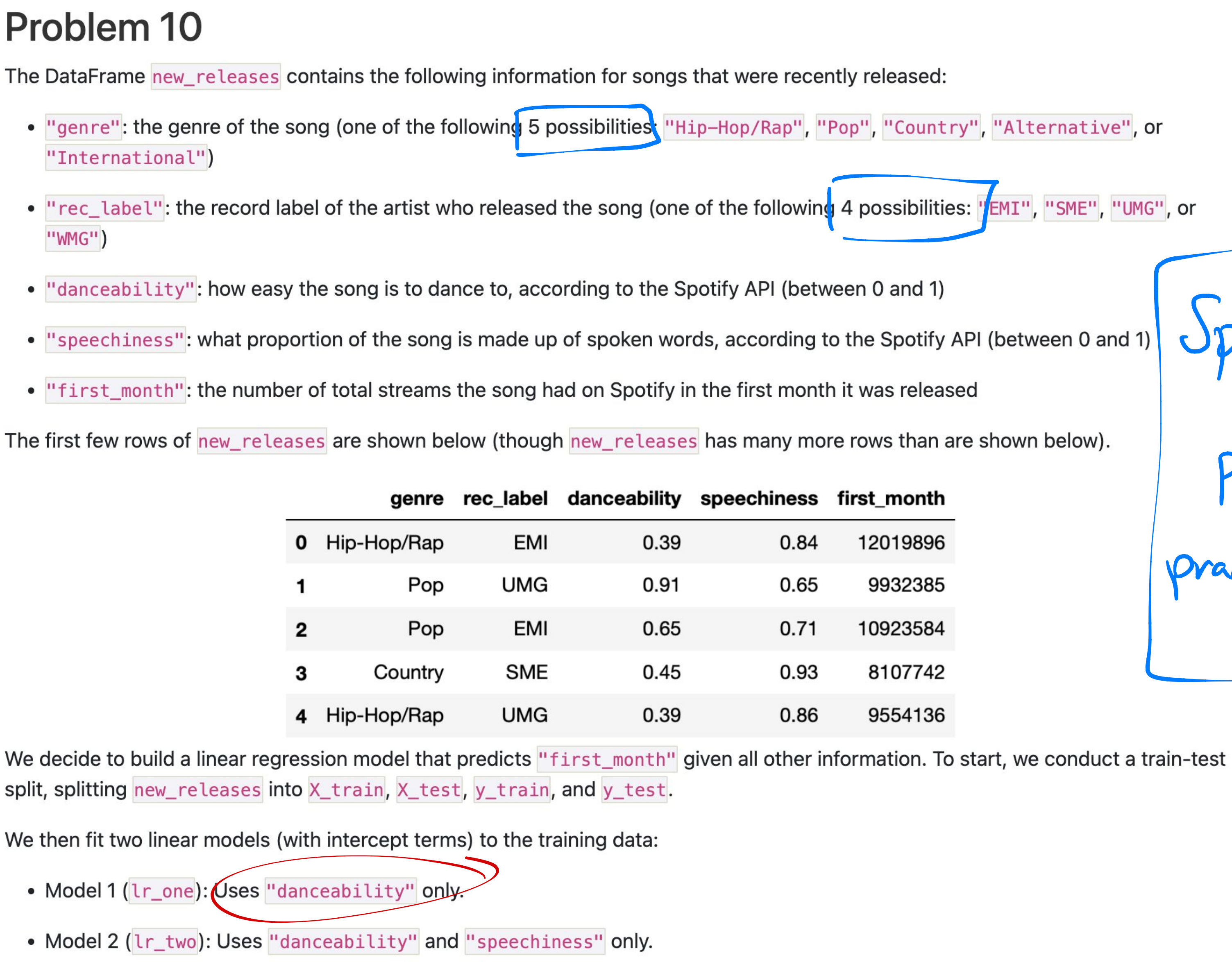
- Project 4 is due on **Thursday, March 21st**.
  - **No slip days allowed!**
- The Final Exam is on **Tuesday, March 19th from 3-6PM in Pepper Canyon Hall 109 (the same room as lecture)**.
  - Practice by working through old exams at practice.dsc80.com. The Spring 2023 Final Exam was recently added.
  - You can bring two double-sided notes sheets that you handwrite.
  - A logistics post on Ed and **assigned seats** are coming soon.
- If at least 80% of the class fills out both **SETs** and the DSC 80-specific **End-of-Quarter Survey** by **Saturday at 8AM**, then the entire class will have **1% of extra credit added to their overall grade**. We value your feedback!
  - As of this morning, the End-of-Quarter Survey had around a 57% completion rate.

Agenda

- We'll work through selected problems from past Final Exams.
- We won't write any code, since you can't run code during the exam. Instead, we'll try to think like the computer ourselves.
- These annotated slides will be posted after lecture is over, as will the ~~solutions to the entire exam~~.
- **Try the problems with us!**
- Towards the end, I'll share some parting thoughts, too.

Spring 2022 Final Exam, Problem 10

Read the problem here.



Problem 10

The DataFrame `new_releases` contains the following information for songs that were recently released:

- `"genre"`: the genre of the song (one of the following 5 possibilities: `"Hip-Hop/Rap"`, `"Pop"`, `"Country"`, `"Alternative"`, or `"International"`)
- `"rec_label"`: the record label of the artist who released the song (one of the following 4 possibilities: `"EMI"`, `"SME"`, `"UMG"`, or `"WMG"`)
- `"danceability"`: how easy the song is to dance to, according to the Spotify API (between 0 and 1)
- `"speechiness"`: what proportion of the song is made up of spoken words, according to the Spotify API (between 0 and 1)
- `"first_month"`: the number of total streams the song had on Spotify in the first month it was released

The first few rows of `new_releases` are shown below (though `new_releases` has many more rows than are shown below).

	genre	rec_label	danceability	speechiness	first_month
0	Hip-Hop/Rap	EMI	0.39	0.84	12019896
1	Pop	UMG	0.91	0.65	9932385
2	Pop	EMI	0.65	0.71	10923584
3	Country	SME	0.45	0.93	8107742
4	Hip-Hop/Rap	UMG	0.39	0.86	9554136

We decide to build a linear regression model that predicts `"first_month"` given all other information. To start, we conduct a train-test split, splitting `new_releases` into `X_train`, `X_test`, `y_train`, and `y_test`.

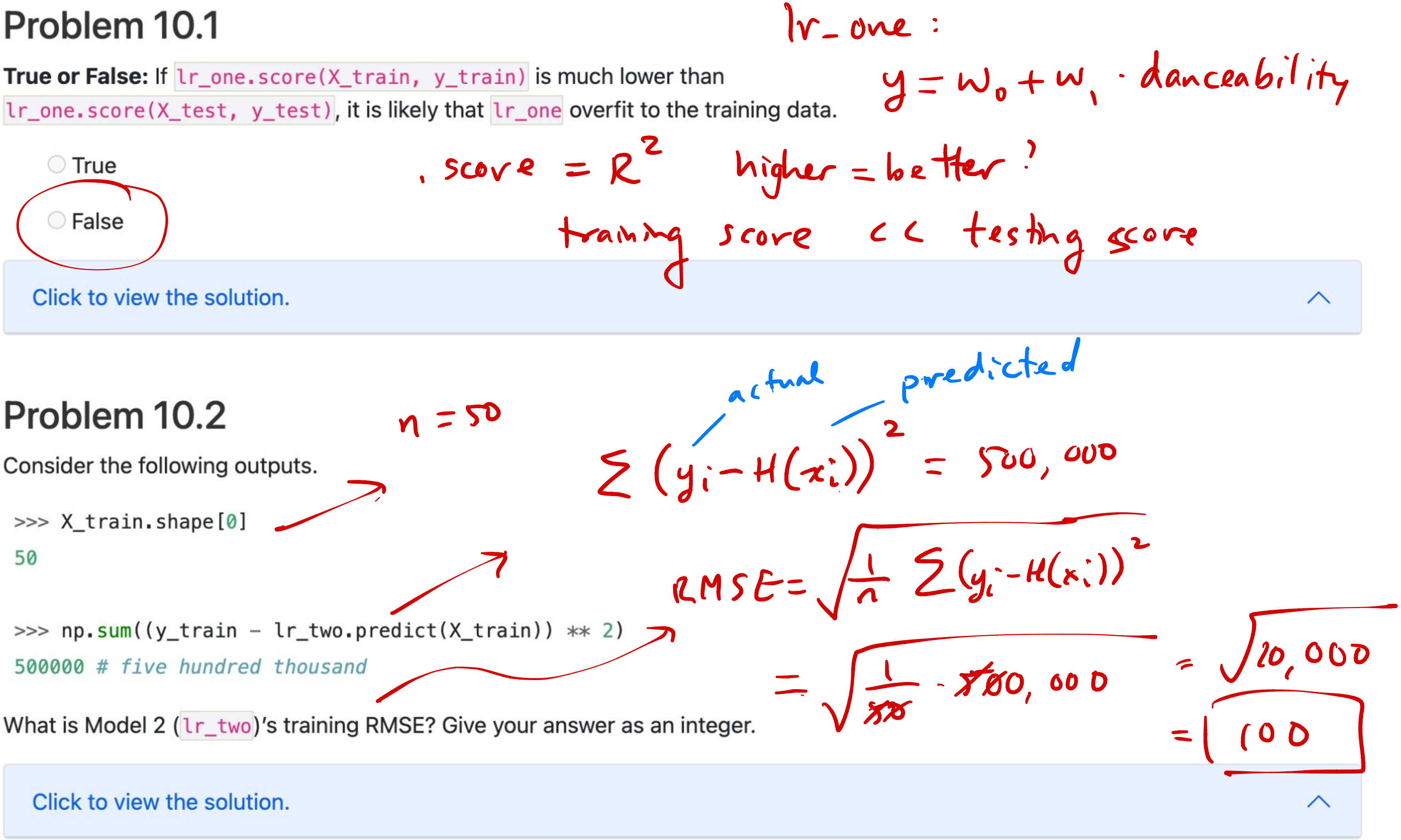
We then fit two linear models (with intercept terms) to the training data:

- Model 1 (`lr_one`): Uses `"danceability"` only.
- Model 2 (`lr_two`): Uses `"danceability"` and `"speechiness"` only.

Spring 22
Final

Problem 10

practice.dsc80.com



Problem 10.1

True or False: If `lr_one.score(X_train, y_train)` is much lower than `lr_one.score(X_test, y_test)`, it is likely that `lr_one` overfit to the training data.

○ True

[✓ ○ False]

lr_one :

$y = w_0 + w_1 \cdot \text{danceability}$

, score = R^2 higher = better ?

training score << testing score

Click to view the solution. ^

Problem 10.2

Consider the following outputs.

n = 50

```
>>> X_train.shape[0]
50

>>> np.sum((y_train - lr_two.predict(X_train)) ** 2)
500000 # five hundred thousand
```

actual predicted

$$\sum (y_i - H(x_i))^2 = 500{,}000$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum (y_i - H(x_i))^2}$$

$$= \sqrt{\frac{1}{\cancel{50}} \cdot \cancel{50}0{,}000} = \sqrt{20{,}000} = \boxed{100}$$

What is Model 2 (`lr_two`)'s training RMSE? Give your answer as an integer.

Click to view the solution. ^

Now, suppose we fit two more linear models (with intercept terms) to the training data:

5 possible   4 possible

- Model 3 (**`lr_drop`**): Uses `"danceability"` and `"speechiness"` as-is, and one-hot encodes `"genre"` and `"rec_label"`, using `OneHotEncoder(drop="first")`. why?
- Model 4 (**`lr_no_drop`**): Uses `"danceability"` and `"speechiness"` as-is, and one-hot encodes `"genre"` and `"rec_label"`, using `OneHotEncoder()`.

Note that the only difference between Model 3 and Model 4 is the fact that Model 3 uses `drop="first"`.

Problem 10.3

How many **one-hot encoded** columns are used in each model? In other words, how many **binary** columns are used in each model? Give both answers as integers.

Hint: Make sure to look closely at the description of `new_releases` at the top of the previous page, and don't include the already-quantitative features.

number of one-hot encoded columns in Model 3 (`lr_drop`) = $(5-1)+(4-1) = 4+3 = \boxed{7}$

number of one-hot encoded columns in Model 4 (`lr_no_drop`) = $5+4 = \boxed{9}$ is_HipHop is_Pop ....

Click to view the solution. ︿

Problem 10.4

Fill in the blank:

`lr_drop.score(X_test, y_test)` is ______

`lr_no_drop.score(X_test, y_test)`.

○ likely greater than

[✓ ○ roughly equal to]

○ likely less than

Click to view the solution.

Model 3's performance $\underset{}{\overset{=}{\rule{3cm}{0.4pt}}}$ Model 4's perf

$$\text{Daisy: streams} = w_0 + w_1 \cdot d_{da} + w_2 \cdot \text{speechiness} + w_3 \cdot \cancelto{1}{(\text{is\_UMG})} + 2000 \cdot \cancelto{1}{(\text{is\_Pop})}$$

equal!

$$\text{Billy: streams} = w_0 + w_1 \cdot d_{bi} + w_2 \cdot \text{speechiness} + w_3 \cdot \cancelto{1}{(\text{is\_UMG})} + 1000 \cdot \cancelto{1}{(\text{is\_Country})}$$

Problem 10.5

Recall, in Model 4 (`lr_no_drop`) we one-hot encoded `"genre"` and `"rec_label"`, and did not use `drop="first"` when instantiating our `OneHotEncoder`.

Suppose we are given the following coefficients in Model 4:

- The coefficient on `"genre_Pop"` is 2000.
- The coefficient on `"genre_Country"` is 1000.
- The coefficient on `"danceability"` is $10^6 = 1{,}000{,}000$.

$$\frac{2000 - 1000}{10^6} = \frac{1}{10^3}$$

Daisy and Billy are two artists signed to the same `"rec_label"` who each just released a new song with the same `"speechiness"`. Daisy is a `"Pop"` artist while Billy is a `"Country"` artist.

Model 4 predicted that Daisy's song and Billy's song will have the same `"first_month"` streams. What is the **absolute difference** between Daisy's song's `"danceability"` and Billy's song's `"danceability"`? Give your answer as a simplified fraction.

d_{da} : Daisy's danceability

d_{bi} : Billy's danceability

Click to view the solution.

Daisy: $\text{streams} = w_0 + w_1 \cdot d_{da} + w_2 \cdot \text{speechness} + w_3 \cdot \cancelto{1}{(\text{is\_UMG})} + 2000 \cdot \cancelto{1}{(\text{is\_Pop})}$

equal!

Billy: $\text{streams} = w_0 + w_1 \cdot d_{bi} + w_2 \cdot \text{speechiness} + w_3 \cdot \cancelto{1}{(\text{is\_UMG})} + 1000 \cdot \cancelto{1}{(\text{is\_Country})}$

$$\cancel{w_0} + w_1 \cdot d_{da} + \cancel{w_2 \cdot \text{speechiness}} + 2000 = \cancel{w_0} + w_1 \cdot d_{bi} + \cancel{w_2 \cdot \text{speechiness}} + 1000$$

$$w_1 \cdot d_{da} - w_1 \cdot d_{bi} = -1000$$

$$w_1 (d_{da} - d_{bi}) = -1000$$

coef on danceability $= 10^6$

$$\Rightarrow |d_{da} - d_{bi}| = \left|\frac{-1000}{10^6}\right| = \left|-\frac{1}{10^3}\right| = \left|-\frac{1}{1000}\right| = \boxed{\frac{1}{1000}}$$

Winter 2023 Final Exam, Problem 7

Read the problem here.

Problem 7

Winter 2023 Final
Problem 7

We decide to build a classifier that takes in a state's demographic information and predicts whether, in a given year:

- The state's mean math score was greater than its mean verbal score (1), or
- the state's mean math score was less than or equal to its mean verbal score (0).

Problem 7.1

1: Math > Verbal
0: Math ≤ Verbal

The simplest possible classifier we could build is one that predicts the same label (1 or 0) every time, independent of all other features.

Consider the following statement:

If `a > b`, then the constant classifier that maximizes training accuracy predicts 1 every time; otherwise, it predicts 0 every time.

For which combination of `a` and `b` is the above statement **not guaranteed** to be true?

Note: Treat `sat` as our training set.

state	Math	Verbal
California	630	750
Washington	552	651
Michigan	715	445
⋮	⋮	⋮

~~Option 1~~: works correctly

```
a = (sat['Math'] > sat['Verbal']).mean()
b = 0.5
```

[✓ Option 2:]

```
a = (sat['Math'] - sat['Verbal']).mean()
b = 0
```

~~Option 3~~: same as Option 1!

```
a = (sat['Math'] - sat['Verbal'] > 0).mean()
b = 0.5
```

~~Option 4~~: works similarly

```
a = ((sat['Math'] / sat['Verbal']) > 1).mean() - 0.5
b = 0
```

options that work satisfy this cond:
$a > b$ implies at least 50% of states have Math > Verbal

Problem 7.2

Suppose we train a classifier, named Classifier 1, and it achieves an accuracy of $\frac{5}{9}$ on our training set.

Typically, root mean squared error (RMSE) is used as a performance metric for regression models, but mathematically, nothing is stopping us from using it as a performance metric for classification models as well.

What is the RMSE of Classifier 1 on our training set? Give your answer as a **simplified fraction**.

classifier's accuracy $= \frac{5}{9}$, RMSE

5/9's of the time, actual = pred,

4/9's of the time,

$$(\text{actual} - \text{pred})^2 = (1-0)^2$$
$$= (0-1)^2$$
$$= 1$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\text{actual} - \text{pred})^2}$$

$$= \sqrt{\frac{5}{9} \cdot 0 + \frac{4}{9} \cdot 1}$$

$$= \sqrt{\frac{4}{9}} = \boxed{\frac{2}{3}}$$

Problem 7.3

While Classifier 1's accuracy on our training set is $\frac{5}{9}$, its accuracy on our test set is $\frac{1}{4}$. Which of the following scenarios is most likely?

- ○ Classifier 1 overfit to our training set; we need to increase its complexity.
- [✓ ○ Classifier 1 overfit to our training set; we need to decrease its complexity.]
- ○ Classifier 1 underfit to our training set; we need to increase its complexity.
- ○ Classifier 1 underfit to our training set; we need to decrease its complexity.

training accuracy >> test accuracy

For the remainder of this question, suppose we train another classifier, named Classifier 2, again on our training set. Its performance on the training set is described in the confusion matrix below. Note that the columns of the confusion matrix have been separately normalized so that each has a sum of 1.

	Actually 0	Actually 1
Predicted 0	0.9	0.4
Predicted 1	0.1	0.6

Problem 7.4

Suppose `conf` is the DataFrame above. Which of the following evaluates to a Series of length 2 whose only unique value is the number `1`?

[✓ ○ `conf.sum(axis=0)`]

○ `conf.sum(axis=1)` $\rightarrow \begin{bmatrix}1.3\\0.7\end{bmatrix}$

pre-norm:

	actually 0	actually 1
pred 0	TN	FN
pred 1	FP	TP

Problem 7.5

Fill in the blank: the ___ of Classifier 2 is guaranteed to be 0.6.

norm: →

		$\frac{TP}{TP+FN}$

○ precision

[✓ ○ recall]

$$\alpha = \frac{A}{A+B}$$

$$1-\alpha = \frac{B}{A+B}$$

	Actually 0	Actually 1
Predicted 0	0.9	0.4
Predicted 1	0.1	0.6

using hint ⇒

0.9B	0.4A
0.1B	0.6A

unnormalized

Problem 7.6

Suppose a fraction α of the labels in the training set are actually 1 and the remaining $1-\alpha$ are actually 0. The accuracy of Classifier 2 is 0.65. What is the value of α?

Hint: If you're unsure on how to proceed, here are some guiding questions:

- Suppose the number of y-values that are actually 1 is A and that the number of y-values that are actually 0 is B. In terms of A and B, what is the accuracy of Classifier 2? Remember, you'll need to refer to the numbers in the confusion matrix above.
- What is the relationship between A, B, and α? How does it simplify your calculation for the accuracy in the previous step?

pre-norm:

	actually 0	actually 1
pred 0	TN	FN
pred 1	FP	TP

norm:
⇒

	$\frac{TP}{TP+FN}$

$$acc = \frac{0.6A + 0.9B}{0.6A + 0.4A + 0.9B + 0.1B} = \frac{0.6A + 0.9B}{A+B}$$

$$= 0.6\alpha + 0.9(1-\alpha) = 0.65$$

$$0.6\alpha + 0.9 - 0.9\alpha = 0.65$$

$$\Rightarrow 0.3\alpha = 0.25$$

$$\alpha = \frac{25}{30} = \boxed{\frac{5}{6}}$$

Winter 2023 Final Exam, Problem 8

Read the problem here.

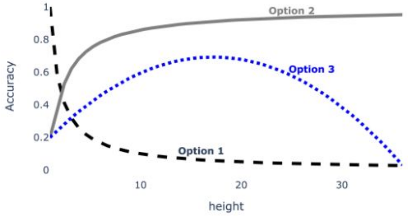
Problem 8.1

`ChickenClassifier`s have many hyperparameters, one of which is `height`. As we increase the value of `height`, the model variance of the resulting `ChickenClassifier` also increases.

First, we consider the training and testing accuracy of a `ChickenClassifier` trained using various values of `height`. Consider the plot below.

Problem 8
Winter 2023
Final

height ↑, complexity ↑



Which of the following depicts **training accuracy vs. `height`**?

○ Option 1

[✓ ○ Option 2]

○ Option 3

as height ↑, complexity ↑,
start overfitting

Which of the following depicts **testing accuracy vs. `height`**?

○ Option 1

○ Option 2

[✓ ○ Option 3]

ChickenClassifiers have another hyperparameter, `color`, for which there are four possible values: `"yellow"`, `"brown"`, `"red"`, and `"orange"`. To find the optimal value of `color`, we perform k-fold cross-validation with $k = 4$. The results are given in the table below.

	Fold 1	Fold 2	Fold 3	Fold 4	row mean
yellow	0.56	0.59	0.39	0.76	0.575
brown	0.42	0.52	0.65	0.48	0.5175
red	0.49	0.51	0.66	0.83	0.6225
orange	0.6	0.49	0.65	0.54	0.57
column mean	0.5175	0.5275	0.5875	0.6525	

Problem 8.2

Which value of `color` has the best average validation accuracy?

○ `"yellow"`

○ `"brown"`

[✓ ○ `"red"`] → highest row mean!

○ `"orange"`

Problem 8.3

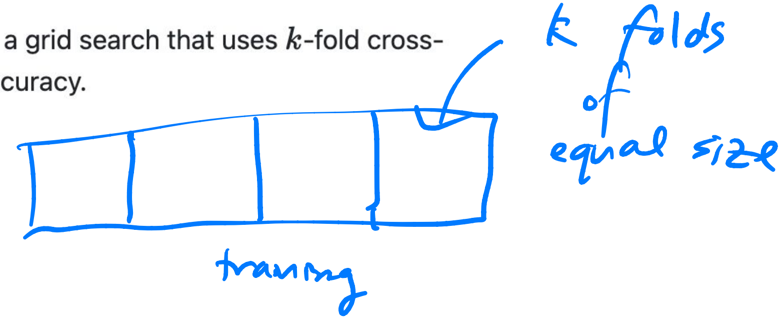
True or False: It is possible for a hyperparameter value to have the best average validation accuracy across all folds, but not have the best validation accuracy in any one particular fold.

[✓ ○ True]

○ False

Problem 8.4

Now, instead of finding the best `height` and best `color` individually, we decide to perform a grid search that uses k-fold cross-validation to find the combination of `height` and `color` with the best average validation accuracy.



For the purposes of this question, assume that:

- We are performing k-fold cross validation.
- Our training set contains n rows, where n is greater than 5 and is a multiple of k.
- There are h_1 possible values of `height` and h_2 possible values of `color`.

Consider the following three subparts:

- A. What is the size of each fold?
- B. How many times is row 5 in the training set used for training?
- C. How many times is row 5 in the training set used for validation?

Choose from the following options.

○ k

○ $\frac{k}{n}$

[✓ ○ $\frac{n}{k}$] A

○ $\frac{n}{k} \cdot (k-1)$

○ $h_1 h_2 k$

[✓ ○ $h_1 h_2 (k-1)$] B

○ $\frac{n h_1 h_2}{k}$

[✓ ○ None of the above] C

(A)

$$\frac{\text{num rows}}{k} = \boxed{\frac{n}{k}}$$

(B)

k-1 times per comb of hyperparams

comb: $h_1 \cdot h_2$

total = $\boxed{h_1 \cdot h_2 \cdot (k-1)}$

(C)

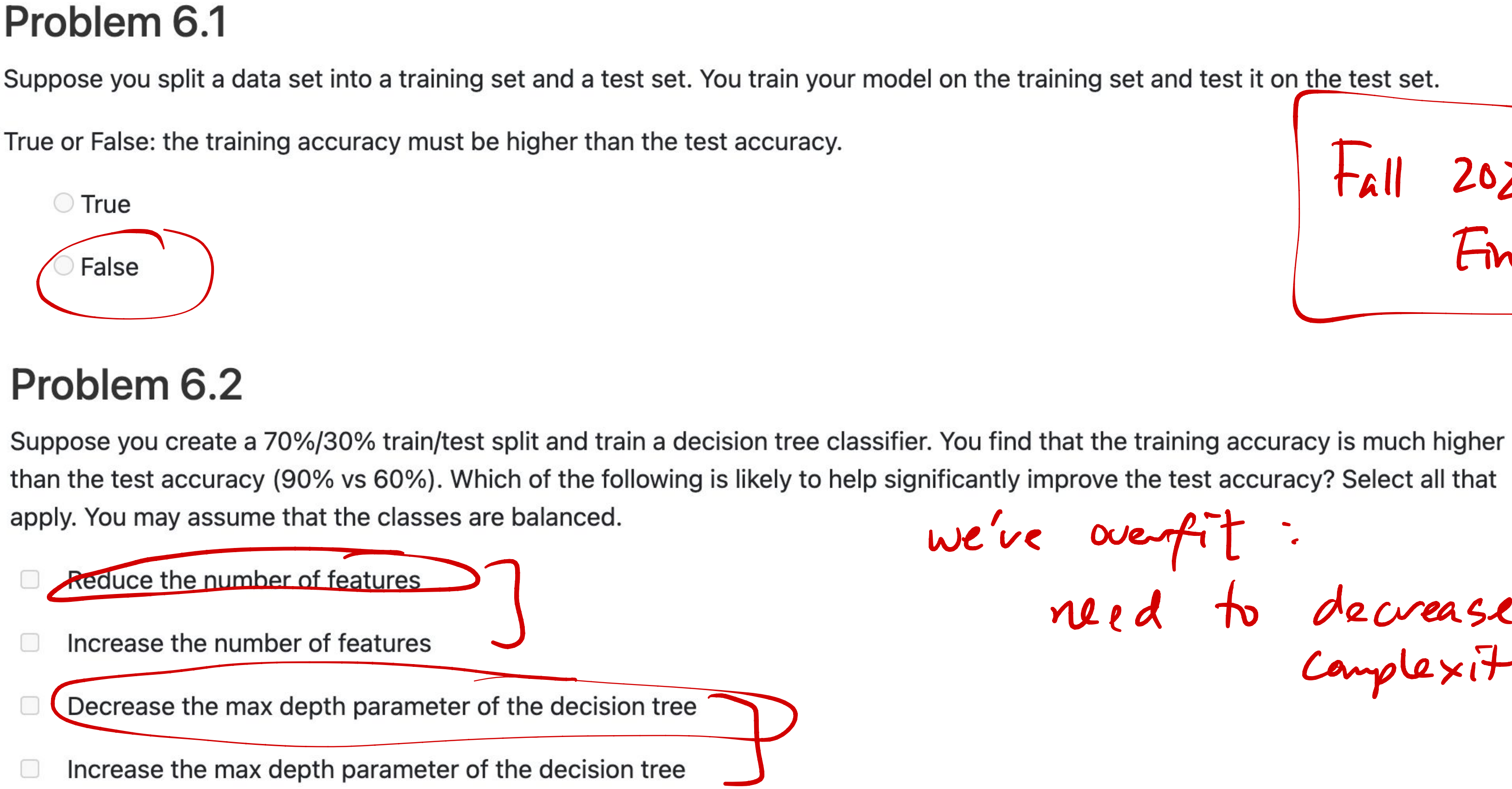
once per comb

total = $h_1 \cdot h_2 \cdot 1 = h_1 \cdot h_2$

$\boxed{\text{None of the above}}$

Fall 2021 Final Exam, Problem 6

Read the problem here.



Problem 6.1

Suppose you split a data set into a training set and a test set. You train your model on the training set and test it on the test set.

True or False: the training accuracy must be higher than the test accuracy.

○ True

[✓ ○ False]

Fall 2021
Final

Problem 6.2

Suppose you create a 70%/30% train/test split and train a decision tree classifier. You find that the training accuracy is much higher than the test accuracy (90% vs 60%). Which of the following is likely to help significantly improve the test accuracy? Select all that apply. You may assume that the classes are balanced.

[✓ ☐ Reduce the number of features]

☐ Increase the number of features

[✓ ☐ Decrease the max depth parameter of the decision tree]

☐ Increase the max depth parameter of the decision tree

we've overfit :
need to decrease
complexity

Problem 6.3

Suppose you are training a decision tree classifier as part of a pipeline with PCA. You will need to choose three parameters: the number of components to use in PCA, the maximum depth of the decision tree, and the minimum number of points needed for a leaf node. You'll do this using sklearn's GridSearchCV which performs a grid search with k-fold cross validation.

Suppose you'll try 3 possibilities for the number of PCA parameters, 5 possibilities for the max depth of the tree, 10 possibilities for the number of points needed for a leaf node, and use k=5 folds for cross-validation.

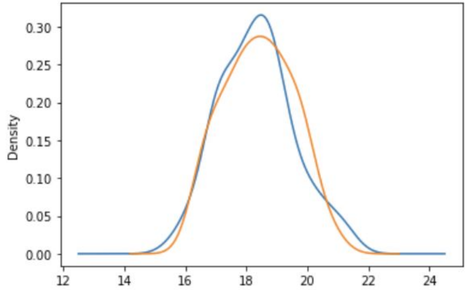
How many times will the model be trained by GridSearchCV?

$3 \cdot 5 \cdot 10 \cdot 5 = \boxed{750}$

k times per combination of hyperparameters

Problem 6.4

The plot below shows the distribution of reported gas mileage for two models of car.



What test statistic is the best choice for testing whether the two empirical distributions came from different underlying distributions?

○ the ~~TVD~~ not categorical!

○ the absolute difference in means

○ the signed difference in means

[✓ ○ the Kolmogorov-Smirnov Statistic]

Problem 6.5

Suppose 1000 people are surveyed. One of the questions asks for the person's age. Upon reviewing the results of the survey, it is discovered that some of the ages are missing – these people did not respond with their age. What is the most likely type of this missingness?

○ Missing At Random

○ Missing Completely At Random

[✓ ○ Not Missing At Random]

○ Missing By Design

missingness likely depends on the ages themselves

Problem 6.6

all data : 550 children = 55% children
450 adults = 45% adults

observed : 400 children = 50% children
400 adults = 50% adult

Consider a data set consisting of the height of 1000 people. The data set contains two columns: height, and whether or not the person is an adult.

Suppose that some of the heights are missing. Among those whose heights are observed there are 400 adults and 400 children; among those whose height is missing, 50 are adults and 150 are children.

If the mean height is computed using only the observed data, which of the following will be true?

○ the mean will be biased low

[✓ ○ the mean will be biased high]

○ the mean will be unbiased

obs : fewer kids, more adults

Problem 6.7

We have built two models which have the following accuracies: Model 1: Train score: 93%, Test score: 67%. Model 2: Train score: 84%, Test score: 80% Which of the following model will you choose to use to make future predictions on unseen data? You may assume that the class labels are balanced.

○ Model 1

[✓ ○ Model 2]

testing performance is higher

Problem 6.8

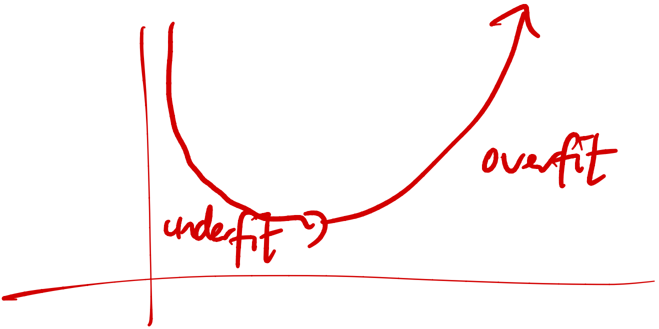
Suppose we retrain a decision tree model, each time increasing the `max_depth` parameter. As we do so, we plot the *test* error. What will we likely see?

[✓ ○ The test error will first decrease, then increase.]

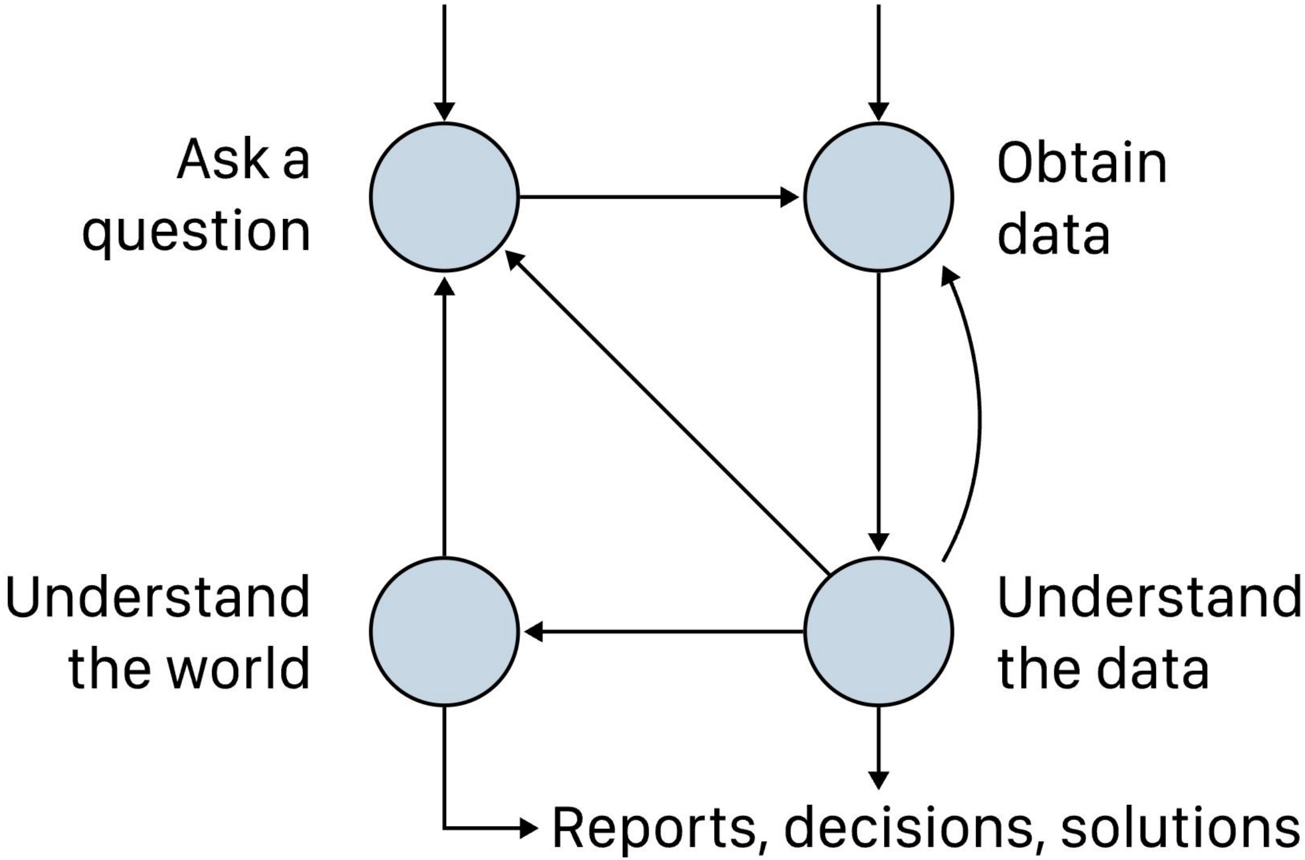
○ The test error will decrease.

○ The test error will first increase, then decrease.

○ The test error will remain unchanged.



Parting thoughts 💭



Ask a question
Obtain data
Understand the world
Understand the data
Reports, decisions, solutions

Course goals ✅

In this course, you...

- **Got a taste of the "life of a data scientist".**
- **Practiced** translating potentially vague questions into quantitative questions about measurable observations.
- **Learned** to reason about 'black-box' processes (e.g. complicated models).
- **Understood** computational and statistical implications of working with data.
- **Learned** to use real data tools (and rely on documentation).

Course outcomes ✅

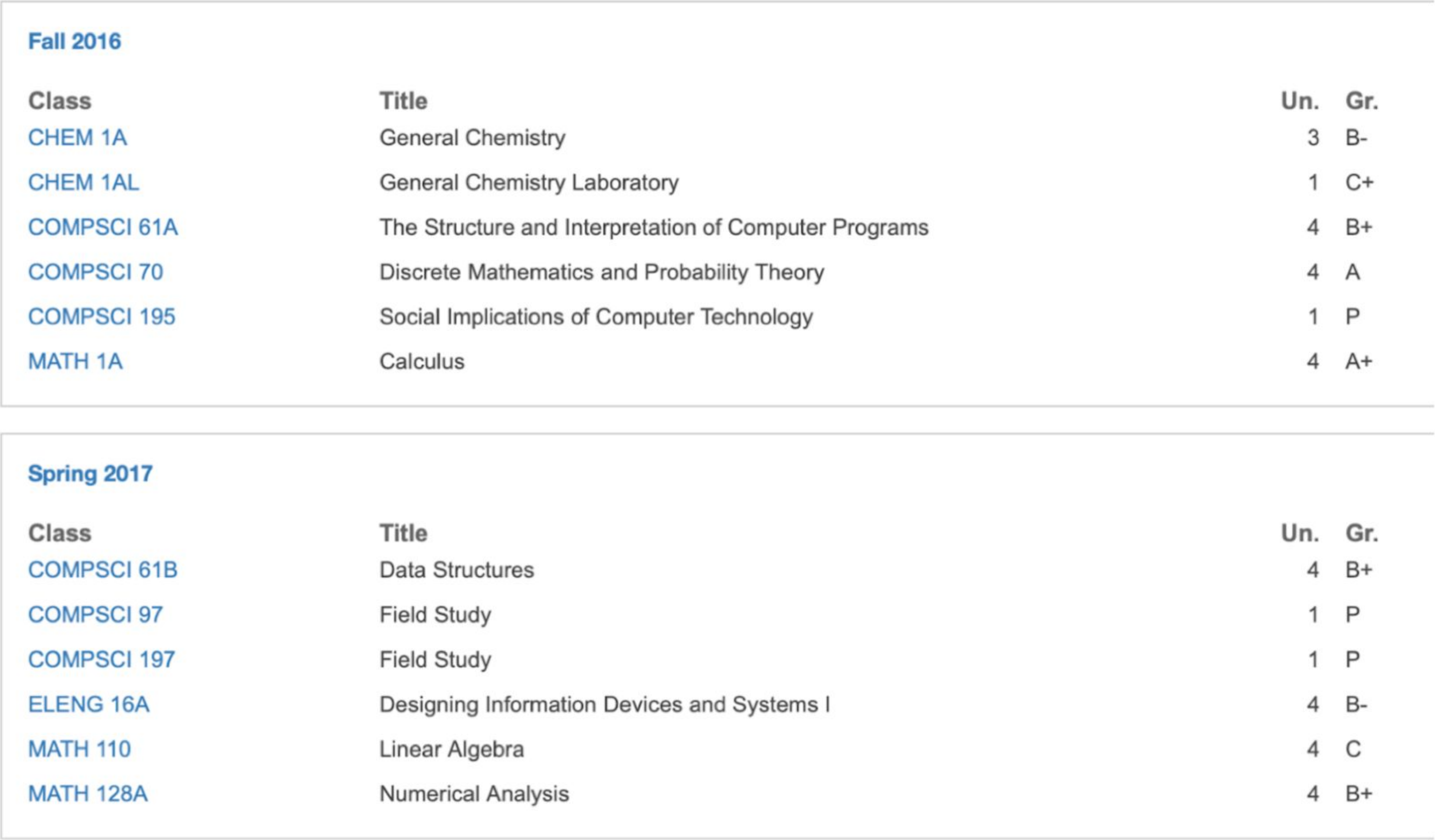
Now, you…

- Are **prepared** for internships and data science "take home" interviews!
- Are **ready** to create your own portfolio of personal projects.
- Have the **background** and **maturity** to succeed in the upper-division.

Topics

We covered **a lot** this quarter! You're now among the most qualified data scientists in the world.

- Week 1: From `babypandas` to `pandas`.
- Week 2: DataFrames.
- Week 3: Working with messy data, hypothesis and permutation testing.
- Week 4: Missing values.
- Week 5: HTML, **Midterm Exam**.
- Week 6: Web and text data.
- Week 7: Text data, modeling.
- Week 8: Feature engineering and generalization.
- Week 9: Modeling in `sklearn`.
- Week 10: Classifier evaluation, fairness, conclusion.
- Week 11: **Final Exam**.



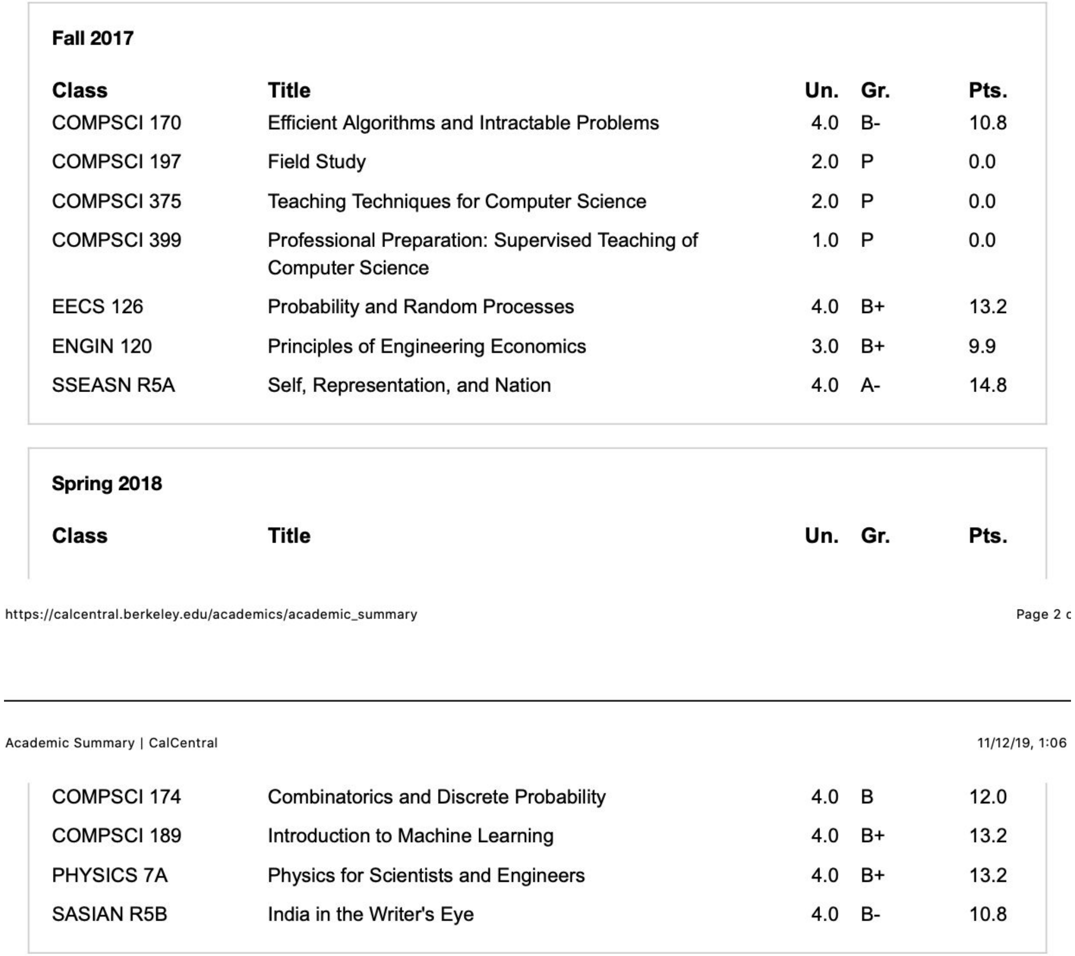
Fall 2016

Class	Title	Un.	Gr.
CHEM 1A	General Chemistry	3	B-
CHEM 1AL	General Chemistry Laboratory	1	C+
COMPSCI 61A	The Structure and Interpretation of Computer Programs	4	B+
COMPSCI 70	Discrete Mathematics and Probability Theory	4	A
COMPSCI 195	Social Implications of Computer Technology	1	P
MATH 1A	Calculus	4	A+

Spring 2017

Class	Title	Un.	Gr.
COMPSCI 61B	Data Structures	4	B+
COMPSCI 97	Field Study	1	P
COMPSCI 197	Field Study	1	P
ELENG 16A	Designing Information Devices and Systems I	4	B-
MATH 110	Linear Algebra	4	C
MATH 128A	Numerical Analysis	4	B+

Suraj’s freshman year transcript.



Fall 2017

Class	Title	Un.	Gr.	Pts.
COMPSCI 170	Efficient Algorithms and Intractable Problems	4.0	B-	10.8
COMPSCI 197	Field Study	2.0	P	0.0
COMPSCI 375	Teaching Techniques for Computer Science	2.0	P	0.0
COMPSCI 399	Professional Preparation: Supervised Teaching of Computer Science	1.0	P	0.0
EECS 126	Probability and Random Processes	4.0	B+	13.2
ENGIN 120	Principles of Engineering Economics	3.0	B+	9.9
SSEASN R5A	Self, Representation, and Nation	4.0	A-	14.8

Spring 2018

Class	Title	Un.	Gr.	Pts.
COMPSCI 174	Combinatorics and Discrete Probability	4.0	B	12.0
COMPSCI 189	Introduction to Machine Learning	4.0	B+	13.2
PHYSICS 7A	Physics for Scientists and Engineers	4.0	B+	13.2
SASIAN R5B	India in the Writer's Eye	4.0	B-	10.8

https://calcentral.berkeley.edu/academics/academic_summary Page 2 c

Academic Summary | CalCentral 11/12/19, 1:06

Suraj's sophomore year transcript.

Thank you!

- This course would not have been possible without our TA and 9 tutors: Dylan Stockard, Aritra Das, Gabriel Cha, Ethan Shapiro, Weiyue Li, Jasmine Lo, Harshi Saha, Yutian Shi, Tiffany Yu, and Diego Zavalza.
- Don’t be a stranger – our contact information is at dsc80.com/staff!
  - This quarter’s course website (and podcasts) will remain online permanently at dsc-courses.github.io.
- Apply to be a tutor in the future! Learn more here.

This could be you!

Good luck on the Final Exam, and enjoy your spring break! 🌴