Lecture 13: Midterm Review

EECS 398: Practical Data Science, Spring 2025

practicaldsc.org • github.com/practicaldsc/sp25 • 📣 See latest announcements [here on Ed]

Agenda

- We'll start by working through the first 8 questions of the Fall 2024 Final Exam: study.practicaldsc.org/fa24-final.
- I'll post these annotated slides after lecture.
- There's another review worksheet available at study.practicaldsc.org/mt-review; from 3:30-5PM, I can answer questions about it (or anything else).

	Type	Brand	Name	Price	Rating	Num Ingredients	Sensitive
0	Eye cream	PERRICONE MD	PRE:EMPT SERIES™ Brightening Eye Cream	55	4.2	33	1
1	Cleanser	CLINIQUE	Pep-Start 2-in-1 Exfoliating Cleanser	19	3.1	36	0
2	Eye cream	PETER THOMAS ROTH	FIRMx™ 360 Eye Renewal	75	5.0	42	0
3	Treatment	KIEHL'S SINCE 1851	Clearly Corrective™ Dark Spot Solution	50	4.5	24	1
4	Cleanser	PETER THOMAS ROTH	Irish Moor Mud Purifying Cleanser Gel	38	3.6	23	0

Problem 1

An expensive product is one that costs **at least $100**.

Problem 1.1

Write an expression that evaluates to the **proportion** of products in `skin` that are expensive.

	Type	Brand	Name	Price	Rating	Num Ingredients	Sensitive
0	Eye cream	PERRICONE MD	PRE:EMPT SERIES™ Brightening Eye Cream	55	4.2	33	1
1	Cleanser	CLINIQUE	Pep-Start 2-in-1 Exfoliating Cleanser	19	3.1	36	0
2	Eye cream	PETER THOMAS ROTH	FIRMx™ 360 Eye Renewal	75	5.0	42	0
3	Treatment	KIEHL'S SINCE 1851	Clearly Corrective™ Dark Spot Solution	50	4.5	24	1
4	Cleanser	PETER THOMAS ROTH	Irish Moor Mud Purifying Cleanser Gel	38	3.6	23	0

Problem 1.2

Fill in the blanks so that the expression below evaluates to the number of brands that sell **fewer than 5** expensive products.

```
skin.groupby(__(i)__).__(ii)__(__(iii)__)["Brand"].nunique()
```

`(i)`:

- ○ `"Brand"`
- ○ `"Name"`
- ○ `"Price"`
- ○ `["Brand", "Price"]`

`(ii)`:

- ○ `agg`
- ○ `count`
- ○ `filter`
- ○ `value_counts`

`(iii)`: (Free response)

Problem 3

Consider the Series `small_prices` and `vc`, both of which are defined below.

```
small_prices = pd.Series([36, 36, 18, 100, 18, 36, 1, 1, 1, 36])

vc = small_prices.value_counts().sort_values(ascending=False)
```

In each of the parts below, select the value that the provided expression evaluates to. If the expression errors, select "Error".

○ 0	`vc.iloc[0]`
○ 1	`vc.loc[0]`
○ 2	`vc.index[0]`
○ 3	`vc.iloc[1]`
○ 4	`vc.loc[1]`
○ 18	`vc.index[1]`
○ 36	
○ 100	
○ Error	
○ None of these	

Problem 4

Consider the DataFrames `type_pivot`, `clinique`, `fresh`, and `boscia`, defined below.

```
type_pivot = skin.pivot_table(index="Type",
                              columns="Brand",
                              values="Sensitive",
                              aggfunc=lambda s: s.shape[0] + 1)
```

```
clinique = skin[skin["Brand"] == "clinique"]
fresh = skin[skin["Brand"] == "fresh"]
boscia = skin[skin["Brand"] == "BOSCIA"]
```

Three columns of `type_pivot` are shown below **in their entirety**.

Brand	CLINIQUE	FRESH	BOSCIA
Type			
Cleanser	6.0	NaN	2.0
Eye cream	4.0	NaN	2.0
Face Mask	3.0	4.0	4.0
Moisturizer	3.0	3.0	NaN
Sun protect	2.0	NaN	NaN

In each of the parts below, give your answer as an **integer**.

Problem 4.1

How many rows are in the following DataFrame?

```
clinique.merge(fresh, on="Type", how="inner")
```

Problem 4.2

How many rows are in the following DataFrame?

```
(clinique.merge(fresh, on="Type", how="outer")
         .merge(boscia, on="Type", how="outer"))
```

Problem 5

Consider a sample of 60 skincare products. The name of one product from the sample is given below:

“our **drops** cream is the best **drops drops** for eye **drops drops** proven formula..."

The total number of terms in the product name above is unknown, but we know that the term **drops** only appears in the name 5 times.

Suppose the TF-IDF of **drops** in the product name above is $\frac{2}{3}$. Which of the following statements are **NOT possible**, assuming we use a base-2 logarithm? Select all that apply.

- [ ] **All** 60 product names contain the term **drops**, **including** the one above.
- [ ] 14 **other** product names contain the term **drops**, in addition to the one above.
- [ ] None of the 59 **other** product names contain the term **drops**.
- [ ] There are 15 terms in the product name above **in total**.
- [ ] There are 25 terms in the product name above **in total**.

Problem 6

Suppose `soup` is a BeautifulSoup object representing the homepage of a Sephora competitor.

Furthermore, suppose `prods`, defined below, is a list of strings containing the name of every product on the site.

```
prods = [row.get("prod") for row in soup.find_all("row", class_="thing")]
```

Given that `prods[1]` evaluates to `"Cleansifier"`, which of the following options describes the source code of the site?

- Option 1:

```
<row class="thing">prod: Facial Treatment Essence</row>
<row class="thing">prod: Cleansifier</row>
<row class="thing">prod: Self Tan Dry Oil SPF 50</row>
...
```

- Option 2:

```
<row class="thing" prod="Facial Treatment Essence"></row>
<row class="thing" prod="Cleansifier"></row>
<row class="thing" prod="Self Tan Dry Oil SPF 50"></row>
...
```

- Option 3:

```
<row prod="thing" class="Facial Treatment Essence"></row>
<row prod="thing" class="Cleansifier"></row>
<row prod="thing" class="Self Tan Dry Oil SPF 50"></row>
...
```

- Option 4:

```
<row class="thing">prod="Facial Treatment Essence"</row>
<row class="thing">prod="Cleansifier"</row>
<row class="thing">prod="Self Tan Dry Oil SPF 50"</row>
...
```

Problem 7

Consider a dataset of n values, $y_1, y_2, \ldots, y_n$, all of which are **positive**. We want to fit a constant model, $H(x) = h$, to the data.

Let h_p^* be the optimal constant prediction that minimizes average degree-p loss, $R_p(h)$, defined below.

$$R_p(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|^p$$

For example, h_2^* is the optimal constant prediction that minimizes $R_2(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|^2$.

In each of the parts below, determine the value of the quantity provided. By "the data", we are referring to $y_1, y_2, \ldots, y_n$.

- ○ The standard deviation of the data
- ○ The variance of the data
- ○ The mean of the data
- ○ The median of the data
- ○ The midrange of the data, $\frac{y_{\min} + y_{\max}}{2}$
- ○ The mode of the data
- ○ None of the above

h_0^*

h_1^*

$R_1(h_1^*)$

h_2^*

$R_2(h_2^*)$

Now, suppose we want to find the optimal constant prediction, h_{U}^*, using the "Ulta" loss function, defined below.

$$L_U(y_i, h) = y_i(y_i - h)^2$$

Problem 7.6

To find h_{U}^*, suppose we minimize average Ulta loss (with no regularization). How does h_{U}^* compare to the mean of the data, M?

- ○ $h_{\text{U}}^* > M$
- ○ $h_{\text{U}}^* \geq M$
- ○ $h_{\text{U}}^* = M$
- ○ $h_{\text{U}}^* \leq M$
- ○ $h_{\text{U}}^* < M$

Now, to find the optimal constant prediction, we will instead minimize **regularized** average Ulta loss, $R_\lambda(h)$, where λ is a non-negative regularization hyperparameter:

$$R_\lambda(h) = \left(\frac{1}{n}\sum_{i=1}^{n} y_i(y_i - h)^2\right) + \lambda h^2$$

It can be shown that $\frac{\partial R_\lambda(h)}{\partial h}$, the derivative of $R_\lambda(h)$ with respect to h, is:

$$\frac{\partial R_\lambda(h)}{\partial h} = -2\left(\frac{1}{n}\sum_{i=1}^{n} y_i(y_i - h) - \lambda h\right)$$

Problem 7.7

Find h^*, the constant prediction that minimizes $R_\lambda(h)$. Show your work, and put a $\boxed{\text{box}}$ around your final answer, which should be an **expression in terms of** y_i**,** n**, and/or** λ.

Problem 8

Suppose we want to fit a simple linear regression model (using squared loss) that predicts the number of ingredients in a product given its price. We're given that:

- The average cost of a product in our dataset is \$40, i.e. $\bar{x} = 40$.
- The average number of ingredients in a product in our dataset is 15, i.e. $\bar{y} = 15$.

The intercept and slope of the regression line are $w_0^* = 11$ and $w_1^* = \frac{1}{10}$, respectively.

Problem 8.1

Suppose Victors' Veil (a skincare product) costs \$40 and has 11 ingredients. What is the squared loss of our model's predicted number of ingredients for Victors' Veil? Give your answer as a **number**.

Problem 8

Suppose we want to fit a simple linear regression model (using squared loss) that predicts the number of ingredients in a product given its price. We're given that:

- The average cost of a product in our dataset is \$40, i.e. $\bar{x} = 40$.
- The average number of ingredients in a product in our dataset is 15, i.e. $\bar{y} = 15$.

The intercept and slope of the regression line are $w_0^* = 11$ and $w_1^* = \frac{1}{10}$, respectively.

Problem 8.2

Is it possible to answer part (a) above **just** by knowing $\bar{x}$ and $\bar{y}$, i.e. **without** knowing the values of w_0^* and w_1^*?

- ○ Yes; the values of w_0^* and w_1^* don't impact the answer to part (a).
- ○ No; the values of w_0^* and w_1^* are necessary to answer part (a).