Data 100, Spring 2022 **April 1, 2022**

Discussion #10

Regularization and Bias-Variance Tradeoff

1. We will use a simple constant model $f_\theta(x) = \theta$ to show the effects of regularization on bias and variance. For the sake of simplicity, we will assume that there is no noise or observational variance, so the ground truth output is equal to the observed outputs: $g(x) = Y$.

   (a) Recall that the optimal solution for the constant model with an MSE loss and a dataset $\mathcal{D}$ with $y_1, y_2, ..., y_n$ is the mean $\bar{y}$.

   Suppose that we use L-2 regularization with coefficient $\lambda > 0$ for training another constant model. Derive the optimal solution to this new constant model **with L-2 regularization** to minimize the objective function below.

$$\arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2 + \lambda\theta^2$$

   (b) Using the bias-variance decomposition, show that the optimal constant model's expected loss on a sample test point (x, Y) in terms of the training data y is equal to the following.

$$\mathbb{E}_\mathcal{D}[(Y - f_{\hat{\theta}}(x))^2] = (Y - \mathbb{E}_\mathcal{D}[\bar{y}])^2 + \text{Var}_\mathcal{D}(\bar{y})$$

   Note: the subscript next to the expectation and variance simply lets you know what is random inside the expectation (i.e. what is the expectation taken over?). In this case, we calculate the expectation and variance of $\bar{y}$ across datasets D.

(c) Redo part (b) with the constant model with the optimal constant model with L-2 regularization to derive the expected loss on a sample point. Show that the obtained bias-variance decomposition is the following.

$$\mathbb{E}_{\mathcal{D}}[(Y - f_{\hat{\theta}}(x))^2] = (Y - \frac{1}{1+\lambda}\mathbb{E}_{\mathcal{D}}[\bar{y}])^2 + \frac{1}{(1+\lambda)^2}\text{Var}_{\mathcal{D}}(\bar{y})$$

Hint: Use your result from part (a), along with some properties of expectations and variances!

(d) Remark on how regularization has affected the model bias and model variance as λ increases. Consider what would happen to these quantities as $\lambda \to \infty$.

SQL Syntax

All SQL queries should follow this basic framework. Note that the order of the clauses matter.

```
SELECT [DISTINCT] ___<columns>___
FROM ___<tables>___
[WHERE ___<predicate>___]
[GROUP BY ___<columns>___]
[HAVING ___<predicate>___]
[ORDER BY ___<columns>___]
[LIMIT ___<number of rows>___]
```

2. For this question, we will be working with the UC Berkeley Undergraduate Career Survey dataset, named `survey`. Each year, the UC Berkeley career center surveys graduating seniors for their plans after graduating. Below is a sample of the full dataset. The full dataset contains many thousands of rows.

j_name	**c_name**	**c_location**	**m_name**
Llama Technician	Google	MOUNTAIN VIEW	EECS
Software Engineer	Salesforce	SF	EECS
Open Source Maintainer	Github	SF	Computer Science
Big Data Engineer	Microsoft	REDMOND	Data Science
Data Analyst	Startup	BERKELEY	Data Science
Analyst Intern	Google	SF	Philosophy

Each record of the `survey` table is an entry corresponding to a student. We have the job title, company information, and the student's major.

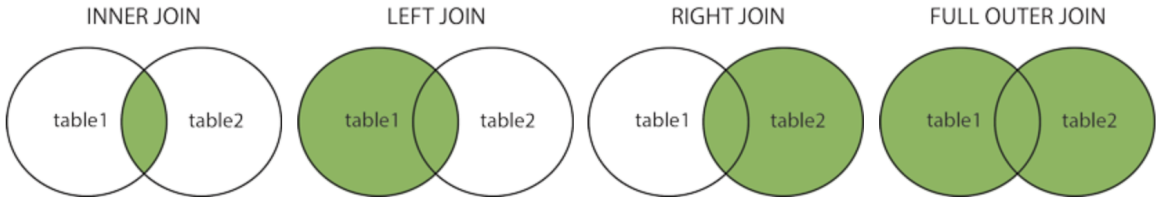
(a) Write a SQL query that selects all data science major graduates that got jobs in Berkeley. The result generated by your query should include all 4 columns.

```
_________________ FROM survey
_____________________________
_____________________________
```

(b) Write a SQL query to find the top 5 popular companies that data science graduates will work at, from most popular to 5th most popular.

```
SELECT c_name, _____________ AS count
FROM survey
WHERE _____________ = "Data Science"
GROUP BY _____________
ORDER BY _____________
LIMIT 5
```

SQL Joins



Note: You do not need the JOIN keyword to join SQL tables. The following are equivalent:

```
SELECT column1, column2
FROM table1, table2
WHERE table1.id = table2.id;
```

```
SELECT column1, column2
FROM table1 JOIN table2
ON table1.id = table2.id;
```

3. In the figure above, assume table1 has m records, while table2 has n records. Describe which records are returned from each type of join. What is the maximum possible number of records returned in each join? Consider the cases where on the joined field, (1) both tables have unique values; and (2) both tables have duplicated values.

4. Consider the following real estate schema:

```
Homes(home_id int, city text, bedrooms int, bathrooms int,
area int)
Transactions(home_id int, buyer_id int, seller_id int,
transaction_date date, sale_price int)
Buyers(buyer_id int, name text)
Sellers(seller_id int, name text)
```

Fill in the blanks in the SQL query to find the id and selling price for each home in Berkeley. If the home has not been sold yet, **the price should be NULL**.

```
SELECT _____
FROM _____
_____ JOIN _____
ON _____
WHERE _____;
```

SQL Queries (Extra)

5. Examine this schema for these two tables:

```
CREATE TABLE owners (
    id integer,
    name text,
    age integer,
    PRIMARY KEY (id)
);
```

```
CREATE TABLE cats (
    id integer,
    owner_id integer,
    name text,
    breed text,
    age integer,
    PRIMARY KEY (id),
    FOREIGN KEY (owner_id) REFERENCES owners
);
```

(a) Write a SQL query to figure out the number of cats, over the age of 10, of each breed of cat.

(b) Write a SQL query to figure out the number of cats each owner owns for owners whose id is greater than 10.

(c) Write a SQL query to figure out the ownerid/owner of the one cat owner who owns the most cats.

(d) Write a SQL query to figure out the names of all of the cat owners who have a cat named Apricot.

(e) It is possible to have a cat with an owner not in the owners table.

◯ A. True ◯ B. False

(f) Write a SQL query to get a random sample of 5 random Siamese Cat (a cat breed) with a name that starts with the letter A.

(g) (Challenge) Write a SQL query to create an almost identical table as cats, except with an additional column 'Nickname' that has the value 'Kitten' for cats less than or equal to the age of 1, 'Catto' for cats between 1 and 15, and 'Wise One' for cats older than or equal to 15.

(h) (Challenge) Write a SQL query to select all rows from the `cats` table that have cats of the top 5 most popular cat breeds.

SQL (Bonus)

6. We wish to convert each Pandas expression to SQL assuming `data` represents a Pandas DataFrame containing 3 columns: `name`, `rank`, and `year`. Both the rank and year are stored as integers.

   (a) `(data['rank'].T.dot(data['year'])) / (data['rank'] ** 2).sum()`

   (b)
   ```
   data.loc[data[‘rank’] < 10, ‘name’] \
       .value_counts() \
       .reset_index()
   ```
   Hint: Remember that `value_counts` returns a sorted output!

   (c)
   ```
   data.merge(data, on = ‘name’) \
       .sort_values(by = ‘name_x’, ascending = False)
   ```

   (d)
   ```
   data.groupby([‘name’, ‘rank’]) \
       .apply(lambda sdf: sdf['year'].max()) \
       .reset_index().head(5)
   ```

   (e)
   ```
   data.groupby(['name', 'year']) \
       .filter(lambda sdf: len(sdf) > 5) \
       .groupby(['name', 'year'])['rank'] \
       .min() \
       .reset_index().head(5)
   ```