Final Practice Questions

Instructions:

- These questions are similar to those in the final exam.
- This document is not representative of the length of the final (it's too long).
- Go back and study the material that is associated with each of the practice questions.

Data Generation and Probability Samples

For each of the following questions select the **single best answer.**

1. A political scientist is interested in answering a question about a country composed of three states with exactly 10000, 20000, and 30000 voting adults. To answer this question, a political survey is administered by randomly sampling 25, 50, and 75 voting adults from each town, respectively. Which sampling plan was used in the survey?

- Cluster sampling
- Stratified Sampling
- Quota sampling
- Snowball sampling

2. A deck with 26 cards labeled A through Z is thoroughly shuffled, and the value of the **third** card in the deck is recorded. What is the probability that we observe the letter C on the third card?

- 1/26
- 3/26
- (25/26) * (24/26) * (1/26)
- None of the above.

3. Suppose Iris visits your store to buy some items. She buys toothpaste for $2.00 with probability 0.5. She buys a toothbrush for $1.00 with probability 0.1. Let the random variable X be the total amount Sam spends. What is E[X]? Show your work in the space provided.

- $1.10
- $1.5
- $3.00
- The toothpaste purchase may not be independent of the toothbrush purchase so we cannot compute this expectation.

4. Suppose we have a coin that lands heads 80% of the time. Let the random variable X be the *proportion* of times the coin lands tails out of 100 flips. What is Var[X]? You must show your work in the space provided.

- 0.8
- 0.16
- 0.04
- 0.0016
- 0.008

Hypothesis Testing

A mysterious undergraduate stops you on your way to class and claims that they have learned to flip any coin such that it lands on heads more often than the 50% you'd expect from random

chance. To demonstrate, they takes a penny from their wallet, flip it 10 times, and get heads nine times and only gets tails once.

1. The null hypothesis is that this was pure random chance, and that the probability of getting heads was 50% for each flip. What is the p-value under the null hypothesis of getting 1 or fewer tails out of 10 flips?

2. Suppose the undergraduate flips the coin 28 more times, and they all end up heads. The resulting p value including all 38 flips under the null hypothesis is approximately $p_b = 10^{-10}$. Which of the following are true? Select all that apply,

    - It is extremely unlikely that the undergraduate just happened to get 37 heads by randomly getting heads on 50/50 coin flips.
    - p_b is the probability that the null hypothesis is true.
    - 1 - p_b is the probability that the undergraduate has the skill to flip any arbitrary coin and get heads.
    - If you flipped a fair coin 38 times, p_b is the chance that you'd get at least 37 heads by random chance.
    - The undergraduate has proven beyond any reasonable doubt that she has the skill to flip any coin to land on heads with high probability.
    - None of the above.

Classification

1. Suppose we train a binary classifier on some dataset. Suppose y is the set of true labels, and yˆ is the set of predicted labels.

y	0	0	0	0	0	1	1	1	1	1
$\hat{y}$	0	1	1	1	1	1	1	0	0	0

Determine each of the following quantities

- The number of true positives
- The number of false negatives
- The precision of our classifier. Write your answer in a simplified fraction.

2. You have a classification data set, where *x* is some value and y is the label for that value:

x	y
2	1
3	0
0	1
1	0

(a) [6 Pts] Suppose that $\phi(x) = [\phi_1 \quad \phi_2 \quad \phi_3]^T = \begin{bmatrix}1 & x & x^2\end{bmatrix}^T$ and our model parameters are $\theta^* = \begin{bmatrix}1 & 0 & -2\end{bmatrix}^T$. For the following parts, leave your answer as an expression (do not numerically evaluate log, e, π, etc).

i. Compute $\hat{\mathbb{P}}(y = 1|x = 0)$.

[4 Pts] Suppose $\phi(x) = \begin{bmatrix}1 & x & x\%2\end{bmatrix}^T$, where % is the modulus operator. Are the data from part a linearly separable with these features? If so, give the equation for a separating plane, e.g. $\phi_2 = 3\phi_3 + 1$. Use 1-indexing, e.g. we have ϕ_1, ϕ_2, and ϕ_3. If not, just write "no".

11. [4 Pts] Suppose we have the dataset below.

x	y
1	1
-1	0

Suppose we have the feature set $\phi(x) = [\phi_1 \quad \phi_2]^T = [1 \quad x]^T$. Suppose we use gradient descent to compute the θ which minimizes the KL divergence under a logistic model without regularization, i.e.

$$\arg\min_{\theta} -\frac{1}{n}\sum_{i=1}^{n}(y_i\phi(x_i)^T + log(\sigma(-\phi(x_i)^T\theta)))$$

Select all that are true regarding the data points and the optimal theta value θ.

☐ A. The data is linearly separable.

☐ B. The optimal θ yields an average cross entropy loss of zero.

☐ C. The optimal θ diverges to $-\infty$

☐ D. The optimal θ diverges to $+\infty$

☐ E. The equation of the line that separates the 2 classes is $\phi_2 = 0$.

☐ F. None of the above.

12. Suppose we have the dataset below.

x	y
-3	1
-1	0
1	0
3	1

Suppose we have the feature set $\phi(x) = \begin{bmatrix} 1 & x^2 \end{bmatrix}^T$. Suppose we use gradient descent to compute the θ which minimizes the KL divergence under a logistic model without regularization, i.e.

$$\arg\min_{\theta} -\frac{1}{n}\sum_{i=1}^{n}(y_i\phi(x_i)^T + log(\sigma(-\phi(x_i)^T\theta)))$$

(a) [3 Pts] Explain in 10 words or fewer why the magnitudes of θ_1 and θ_2 will be very large.

(b) [3 Pts] Will the sign of θ_2 be negative or positive?

- ◯ A. Could be either, it depends on where our gradient descent starts
- ◯ B. Positive
- ◯ C. Negative
- ◯ D. Neither, θ_2 will be zero

(c) [3 Pts] If we use L_1 regularization, which of our θ values would you expect to be zero?

- ◯ A. Neither of them
- ◯ B. θ_1
- ◯ C. θ_2
- ◯ D. Both θ_1 and θ_2

Bias and Variance Trade-off

13. In class, we showed that the expected squared error can be decomposed into several important terms:

$$\mathbb{E}[(Y - f_{\hat{\theta}}(x))^2] = \sigma^2 + (h(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2 + \mathbb{E}[(\mathbb{E}[f_{\hat{\theta}}(x)] - f_{\hat{\theta}}(x))^2].$$

(a) [1 Pt] For which of the following reasons are we taking an expectation? In other words, what are the sources of randomness that we are considering in the derivation of the bias-variance tradeoff?

- ☐ A. We chose arbitrary features when doing feature engineering.
- ☐ B. We drew random samples from some larger population when we built our training set.
- ☐ C. There is some noise in the underlying process that generates our observations Y from our features.
- ☐ D. Our x values could have had missing or erroneous data, e.g. participants misreading a question on a survey.
- ☐ E. None of the Above.

(b) [1.5 Pts] Which of the following do we treat as fixed? Select all that apply.

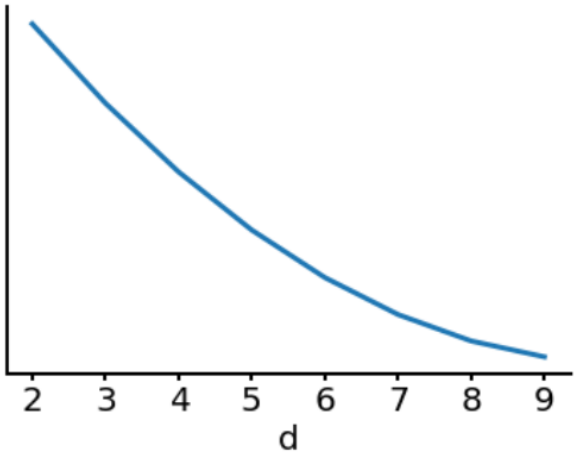
- ☐ A. $\hat{\theta}$
- ☐ B. σ^2
- ☐ C. $h(x)$

(c) [1 Pt] By decreasing model complexity, we are able to decrease σ^2.

- ◯ A. True
- ◯ B. False

14. Your team would like to train a machine learning model in order to predict the next YouTube video that a user will click on based on m features for each of the previous d videos watched by that user. In other words, the total number of features is $m \times d$. You're not sure how many videos to consider.

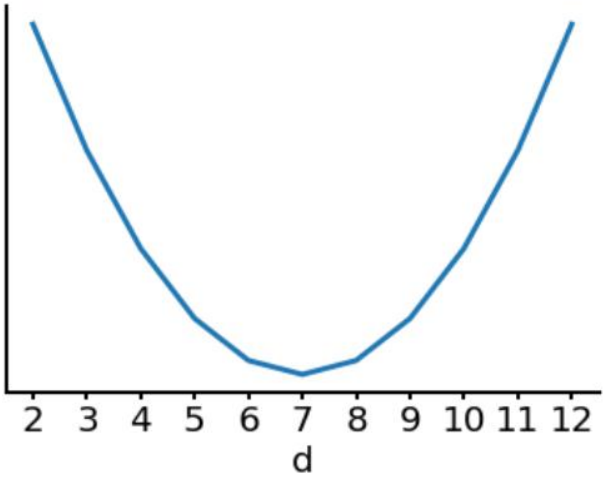
(a) [2 Pts] Your colleague generates the following plot, where the value d is on the x axis. However, they forgot to label the y-axis.



Which of the following could the y axis represent? Select all that apply.

☐ A. Training Error
☐ B. Validation Error
☐ C. Bias
☐ D. Variance

(b) [2 Pts] Your colleague generates the following plot, where the value d is on the x axis. However, they forgot to label the y-axis.



Which of the following could the y axis represent? Select all that apply.

☐ A. Training Error
☐ B. Validation Error
☐ C. Bias
☐ D. Variance

Linear Models

10. Recall from lecture that a linear model is defined as a model where our prediction $\hat{y}$ is given by the equation below, where d is the number of parameters in our model:

$$\hat{y} = f_\theta(x) = \sum_{j=1}^{d} \theta_j \phi_j(x)$$

Which of the following models are linear? **Select all that apply.**

- ☐ A. $f_\theta(x) = \theta_1 x + \theta_2 \sin(x)$
- ☐ B. $f_\theta(x) = \theta_1 x + \theta_2 \sin(x^2)$
- ☐ C. $f_\theta(x) = \theta_1$
- ☐ D. $f_\theta(x) = (\theta_1 x + \theta_2)x$
- ☐ E. $f_\theta(x) = \ln(\theta_1 x + \theta_2) + \theta_3$

11. Suppose we have data about 5 people shown below.

name	level	trials	phase
Magda	1	10	1
Valerie	5	20	-1
Kumar	2	15	1
Octavia	6	30	1
Dorete	6	5	-1

(a) Suppose we want to model the **level** of each person, and use the following constant model: $f_\theta(x) = \theta_1$. What is $\hat{\theta}_1$, the value that minimizes the average L2 loss?

(b) We can also compute $\hat{\theta}$ from the previous part by using the normal equation $\hat{\theta} = (\Phi^T\Phi)^{-1}\Phi^T Y$. If we use the normal equation to compute $\hat{\theta}$, how many rows and columns are in the feature matrix Φ? Write your answer in the form **# rows** $\times$ **# columns**, e.g. 1×1.

(c) What is $(\Phi^T\Phi)^{-1}\Phi^T$? Write your answer in the form of a **Python list**, e.g. $[1, 2, 3]$.

Gradient Descent

12. Momentum is a common variation of gradient descent in which we include the gradient at a previous step of the iteration in our current update equation. More formally it is defined as follows, where γ is the weight of momentum.

$$\theta^{t+1} = \theta^t - \alpha \frac{\partial L}{\partial \theta}\bigg|_{\theta^t} - \gamma \frac{\partial L}{\partial \theta}\bigg|_{\theta^{t-1}}$$

Fill in the code with the following keywords and numbers to implement gradient descent with momentum. Assume when $t = 0$ and $t = -1$, $\theta^t = \texttt{t0}$.

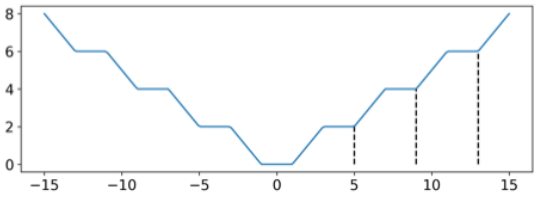
Note that the same keyword/number can be used multiple times; some keywords may not be used at all. Only use one keyword per blank. You may not need all blanks.

theta	phi	y	theta_prev	num_iter	t0
temp	alpha	gamma	range	len	t

```
1  def grad(phi, y, theta):
2      """Returns dL/dtheta. Assume correct implementation."""
3
4  def grad_desc_momentum(phi, y, num_iter, alpha, gamma, t0):
5      """ Returns theta computed after num_iter iterations.
6      phi: matrix, design matrix
7      y: vector, response vector
8      num_iter: scalar, number of iterations to run
9      alpha: scalar, learning rate
10     gamma: scalar, weight of momentum
11     t0: theta for t=0
12     """
13     theta, theta_prev = ___<a>_________, _<b>_____________
14      for _<c>__________ in _<d>___________(_<e>__________):
15          g = grad(phi, y, theta)
16          m = grad(phi, y, _<f>_____________)
17          _<g>_____, _<h>_____ = _<i>_____ - _<j>___ * g - _<k>_____ * m,
18              _<l>_____________
19      return theta
```

Recall that python allows multiple assignment, e.g. "`one, two = two, one`" would swap the values of `one` and `two`.

13. Consider the following function of $f(\theta)$, which alternates between completely flat regions and regions of absolute slope equal to 1. **There is only one correct answer for each part.**



(a) Assuming that θ starts in a flat region that is not a minimum and $\alpha > 0$, will the basic gradient descent algorithm terminate at a minimum? Note that the basic gradient descent algorithm is just the same as version with momentum on the previous page, but where $\gamma = 0$.

◯ A. Never ◯ B. Maybe ◯ C. Yes with enough iterations

(b) Assuming that θ starts in a sloped region and $\alpha > 0$, will the basic gradient descent algorithm find the minimum?

◯ A. Never ◯ B. Maybe ◯ C. Yes with enough iterations

(c) Assuming that θ starts in a flat region that is not a minimum and $\alpha > 0$ and $\gamma > 0$, will the momentum gradient descent algorithm find the minimum?

◯ A. Never ◯ B. Maybe ◯ C. Yes with enough iterations

(d) Assuming that θ starts in a sloped region and $\alpha > 0$ and $\gamma > 0$, will the momentum gradient descent algorithm find the minimum?

◯ A. Never ◯ B. Maybe ◯ C. Yes with enough iterations

(e) Is $f(\theta)$ convex?

◯ A. Yes
◯ B. No
◯ C. No, but $-f(\theta)$ is convex
◯ D. No, but $f(-\theta)$ is convex