Name: ______________________________ NetID: ____________

S&DS 265 / 565

Introductory Machine Learning

Midterm Exam

Tuesday, October 19, 2021

Complete all of the problems. You have 75 minutes to complete the exam.

The exam is closed book, computer, phone, etc. You are allowed one double-sided $8\frac{1}{2} \times 11$ sheet of paper with hand-written notes.

1. *True or False?* (15 points)

   Indicate the *best* answer to each of the following statements.

   The following questions concern bias and variance.

   Suppose we have a data point $Y \sim N(\theta, 4)$, a single random draw from a Normal distribution with mean $\theta = 3$ and variance 4. Consider an estimator $\widehat{\theta}_b = bY$ where $0 \leq b \leq 1$.

TRUE FALSE (1) The squared bias of $\widehat{\theta}_b$ increases with b.

TRUE FALSE (2) The variance of $\widehat{\theta}_b$ increases with b.

TRUE FALSE (3) When $b = \frac{2}{5}$ the squared bias of $\widehat{\theta}_b$ is equal to its variance.

   Recall that for random forests, the parameter m is the number of features included in a random subset selected to split a given node in a tree.

TRUE FALSE (4) As m increases, the squared bias of the ensemble tends to increase.

TRUE FALSE (5) As m increases, the variance of the ensemble tends to increase.

1. *True or False? (continued)*

   The following questions concern logistic regression.

   The logistic regression model with parameters $\beta = (\beta_0, \beta_1, \beta_2)$ and inputs $x = (x_1, x_2)$ defines class probabilities

$$p = \text{logistic}(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2) = \frac{1}{1 + e^{-\beta_0 - \beta_1 \cdot x_1 - \beta_2 \cdot x_2}}$$

   where $p = \mathbb{P}(Y = 1 \,|\, x_1, x_2)$ and $1 - p = \mathbb{P}(Y = 0 \,|\, x_1, x_2)$.

   Circle the correct answer to indicate whether each statement is True or False.

TRUE FALSE (1) If $\beta = (1, 1, 0)$ and $x = (1, 1)$ the model predicts $Y = 1$.

TRUE FALSE (2) If $\beta = (1, 1, 1)$ and $x = (-1, -1)$ the model predicts $Y = 1$.

TRUE FALSE (3) If $\beta = (1, 1, 0)$ and $x = (-1, 1)$, the model assigns equal odds to $Y = 1$ and $Y = 0$.

TRUE FALSE (4) If $\beta = (1, 1, -2)$ and both x_1 and x_2 increase by 1, the model assigns greater probability to $Y = 1$ compared to when x_1 and x_2 are not increased.

TRUE FALSE (5) Logistic regression is a generative model because it assigns a probability to each x.

1. *True or False (continued)*

The following questions concern PCA and dimension reduction.

TRUE FALSE (1) Principal components analysis does not involve prediction accuracy.

TRUE FALSE (2) PCA is a supervised learning algorithm.

TRUE FALSE (3) If our dataset consists of n observations of d different variables, with $d < n$, then the number of possible principal vectors from PCA is n.

TRUE FALSE (4) PCA assumes all predictor variables are quantitative (numerical).

TRUE FALSE (5) The principal components can be used to represent data in a regression model.

2. *Short Answer* (15 points)

   For each of the following five concepts, provide a short definition or explanation. Please limit your response to no more than five sentences.

   (a) *Bias-Variance*. Define squared bias and variance, and explain the bias-variance tradeoff for squared error.

   (b) *Cross-validation*. Define leave-one-out cross validation, and what it is used for.

   (c) *Stochastic gradient descent*. Give the steps for finding the parameters β that minimize a general training loss $\frac{1}{n}\sum_{i=1}^{n} L(y_i, \beta^T x_i)$ using stochastic gradient descent. Express your answer using the derivatives $\ell_j(x, y) \equiv \frac{\partial L(y, \beta^T x)}{\partial \beta_j}$.

2. *Short Answer (continued)*

   (d) *Model complexity*. Give a plot showing how the error, squared bias, and variance of a machine learning model generally vary as a function of the model's complexity.

   (e) *Generative vs. Discriminative*. Explain the key differences between a generative model and a discriminative model for classification. Which approach makes stronger assumptions? Why?

3. *Classification and the Bayes decision rule* (10 points)

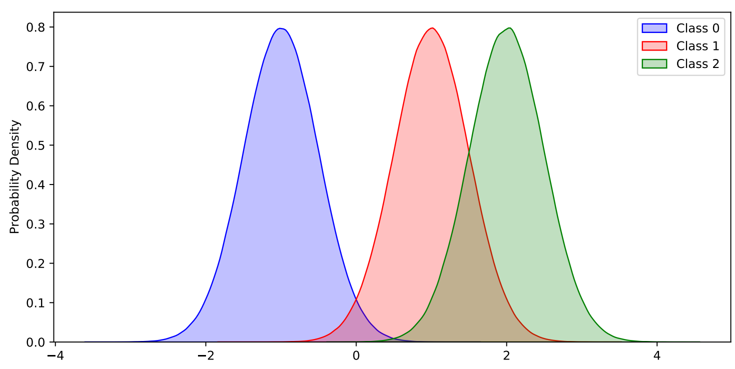
Let X be the predictor variable and y the class variable in a classification model, with $y \in \{0, 1, 2\}$. (For example, think of the Iris flower classification problem.)

(a) Use Bayes' theorem to write $p(y = 0 \,|\, X)$ in terms of the three values $p(y, X)$ for $y \in \{0, 1, 2\}$.

(b) Suppose now that X is a single real number (for example, think of petal width in the Iris flower classification problem). Consider the generative model

$$X \,|\, y = 0 \sim N(-1, \sigma^2) \quad X \,|\, y = 1 \sim N(1, \sigma^2) \quad X \,|\, y = 2 \sim N(2, \sigma^2)$$

where $N(\mu, \sigma^2)$ means a normal distribution with mean μ and variance σ^2. A picture of this is shown below.



Suppose all three classes are equally likely, with $p(y) = \frac{1}{3}$. What is the Bayes decision rule? *Note: Use the plot—only simple arithmetic is needed to answer this question!*

3. *Classification and the Bayes decision rule (continued)*

(c) With the same distributions above, now suppose that $p(y = 0) = p(y = 2) = \frac{1}{2}$. Now what is the Bayes decision rule? *Again: Use the plot—only simple arithmetic is needed to answer this question!*

4. *Coding: Batch gradient descent and prediction* (15 points)

In this problem you are asked to complete the implemention of batch gradient descent and prediction for linear regression.

```
def train_linear_model(X, y, steps=10000, step_size=.001):
  # initialize variables
  beta = # line (a)
  intercept = 0

  # batch gradient descent
  # with fixed step size and number of steps
  for step in np.arange(steps):
      dbeta = # line (b)
      dintercept = # line (c)

      beta = beta - step_size * dbeta
      intercept = intercept - step_size * dintercept
  return beta, intercept


def predict(X, beta, intercept):
  yhat = # line (d)
  return yhat


yhat = predict(Xtest, beta, intercept)
# compute the root mean squared error
rmse = # line (e)
```

Complete the implementation by giving *a single line of Python code* for each of the five missing lines above. Your Python expressions should be syntactically correct; partial credit will be given.

4. *Coding: Batch gradient descent and prediction (continued)*

`line (a):`

`line (b):`

`line (c):`

`line (d):`

`line (e):`