DS5220/CS6140-02

Midterm

October 20, 2017

Time: 1 hour 40min

Name (please print): _____

Show all your work and calculations. Partial credit will be given for work that is partially correct. Points will be deducted for false statements, even if the final answer is correct. Please circle your final answer where appropriate.

This exam is closed-book. You may consult one page with your hand-written notes. Calculators are permitted.

Honor code: I promise not to cheat on this exam. I will neither give nor receive any unauthorized assistance. I will not to share information about the exam with anyone who may be taking it at a different time. I have not been told anything about the exam by someone who has taken it earlier.

Signature: _____ Date: _____

Question	Possible Points	Actual Points
1	20	
2	10	
3	15	
4	20	
5	20	
6	30	
Total	115	

1. For each question below, circle your answer and give a brief (1-2 sentences) explanation.

   (a) **(5 pts)** Zero correlation between any two random variables implies that the two random variables are independent.

   True **False**
   Answer:

   False. Only true for Normal distribution

   (b) **(5 pts)** The slope of simple linear regression with a continuous predictor is the Pearson coefficient of correlation between the predictor and the response.

   True **False**
   Answer:

   False.Only true when the standard deviations of X and Y are the same.

   (c) **(5 pts)** Ridge regularization in linear regression is used for variable selection.

   True **False**
   Answer:

   False. LASSO regression is used for variable selection. Ridge regression regularizes each parameter without setting them to zero.

(d) **(5 pts)** Linear regression is a generalized linear model with identity link function.

True **False**

Answer:

False. Also requires that Y is Normally distributed geven X.

2. A pregnancy test kit has 98.5% true positives, and 0.8% false positives. Suppose a woman using this pregnancy kit is 60% at risk of being pregnant.

(a) **(5 pts)** The test is negative. What is the probability that the woman is pregnant?

Answer:

$$
\begin{aligned}
& P(Pregnant|\,Test\ negative) \\
= \ & \frac{P(Test\ negative|Pregnant)P(Pregnant)}{P(Test\ negative)} \\
= \ & \frac{P(False\ negative|Pregnant)P(Pregnant)}{P(False\ negative|Pregnant)P(Pregnant) + P(True\ negative|Not\ pregnant)P(Not\ pregnant)} \\
= \ & \frac{(1-0.985)\cdot 0.6}{(1-0.985)\cdot 0.6 + (1-0.008)\cdot 0.4} = 0.022
\end{aligned}
$$

(b) **(5 pts)** To double-check, the woman decides to take the test again. This second test, however, turns out to be positive. Assuming that given the true pregnancy status the results of the two tests are independent, what is the probability that the woman is pregnant?

Answer:

$$
\begin{aligned}
& P(Pregnant|\,Test\ positive,\ Test\ negative) \\
= \ & \frac{P(Test\ positive,\ Test\ negative|Pregnant)\cdot P(Pregnant)}{P(Test\ positive,\ Test\ negative)} \\
= \ & \frac{P(Test\ positive|Pregnant)P(Test\ negative|Pregnant)\cdot P(Pregnant)}{P(Test\ positive,\ Test\ negative)} \\
= \ & \frac{0.985\cdot(1-0.985)\cdot 0.6}{0.985\cdot(1-0.985)\cdot 0.6 + 0.008\cdot(1-0.008)\cdot 0.4} = 0.736
\end{aligned}
$$

3. For each question below, circle your answer and give a brief (1-2 sentences) explanation.

(a) **(5 pts)** When we consider a richer space of candidate models, overfitting is more likely.

True **False**
Answer:

True. More predictors more likely cause overfitting

(b) **(5 pts)** When the number of observations in the training set increases to infinity, the model trained on that data will have:

Lower variance **Higher variance** **Same variance**
Answer:

Lower variance. The fitted model is less affected by artifacts of random variation.

(c) **(5 pts)** When the number of observations in the training set increases to infinity, the model trained on that data will have

Lower bias **Higher bias** **Same bias**
Answer:

Same bias. Bias is a property of the model, not of the number of observations. If the model is systematically wrong, it will remain wrong even as the size of the dataset increases.

4. For each question below, circle your answer and give a brief (1-2 sentences) explanation.

(a) **(5 pts)** If we fit logistic regression using gradient descent with the correct step size, we always obtain globally optimal solution.

True **False**
Answer:

True. The objective function of logistic regression is a convex function, which has a global optimum.

(b) **(5 pts)** Stochastic gradient descent has more variability in parameter estimation.

True **False**
Answer:

True. Stochastic gradient descent does not go over all the observations and does not follow the total derivative to update estimates at each step.

(c) **(5 pts)** Stochastic gradient descent performs more computation per update than batch gradient descent.

True **False**
Answer:

False. Only needs computation for one observation, not all observations.

(d) **(5 pts)** Coordinate descent is another word for gradient descent.

True **False**
Answer:

False. Coordinate descent only updates one coordinate at a time.

5. For each question below, circle your answer and give a brief (1-2 sentences) explanation.

(a) **(5 pts)** Nearest neighbors is a parametric method.

True **False**
Answer:

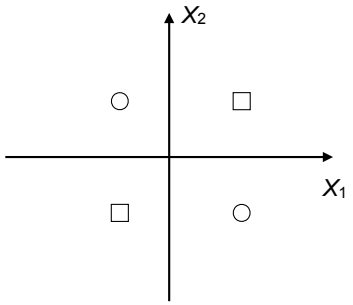
False. KNN does not assume a probability distributions on the samples. It has as many parameters as neighborhoods.

(b) **(5 pts)** To predict the probability of an event with logistic regression, we minimize squared loss.

True **False**
Answer:

False. We minimize the cost function, such as negative log likelihood.

(c) **(5 pts)** In the following dataset, squares correspond to $Y = 1$, and "o" to $Y = 0$. Circle all of the classifiers that will achieve zero error on the training set and explain. (You may circle more than one, or none.)



k-NN, k=1 **k-NN, k=3** **Logistic regression** **LDA** **QDA**
Answer:

None.
KNN, k=1 and k=3: the nearest neighbors are in the wrong class
Logistic regression and LDA have linear decision boundary, yet the dataset is not linearly separable
The dataset does not have enough degrees of freedom to estimate the parameters of QDA and accommodate the boundary of this type

(d) **(5 pts)** LDA with diagonal covariance matrix, where the covariance is regularized towards $\sigma^2 I$, produces the same decision boundary as nearest shrunken centroids.

True **False**

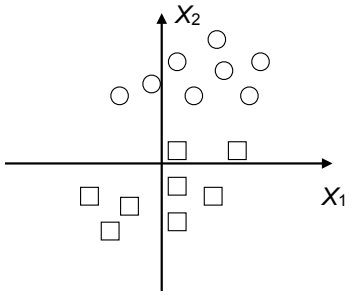
Answer:

False. The methods have different regularization strategy, and make a different use of class priors.

6. In the following dataset, squares correspond to $Y = 1$, and "o" to $Y = 0$. We want to perform binary classification with an additive logistic regression

$$P\{Y = 1\} = \frac{1}{1 + exp\{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)\}}$$

For each question below, circle your answer and give a brief (1-2 sentences) explanation.



(a) **(5 pts)** The parameters of logistic regression estimated from this dataset are unique.

True **False**

Answer:

False. The dataset is nearly separable, and supports multiple linear decision bounds.

(b) To improve the estimation, consider adding regularization

$$log(Likelihood) - \lambda\beta_j^2,\ j = 0, 1, 2$$

In other words, only one of the parameters is regularized in each case.

i. **(5 pts)** When regularizing by β_2 with large λ, the training error

Increases **Decreases** **Stays the same**
Answer:

Increases. When we regularize β_2, the resulting boundary can rely less and less on the value of X_2 and therefore becomes more vertical. For large λ, the training error increases as there is no good linear vertical separator of the training data.

ii. **(5 pts)** When regularizing by β_1 with large λ, the training error

Increases **Decreases** **Stays the same**
Answer:

Remains the same. When we regularize β_1, the resulting boundary can rely less and less on the value of X_1 and therefore becomes more horizontal and the training data can be separated with zero training error with a horizontal linear separator.

iii. **(5 pts)** When regularizing by β_0 with large λ, the training error

Increases **Decreases** **Stays the same**
Answer:

Increases. When we regularize β_0, then the boundary will eventually go through the origin (the intercept set to zero). Based on the figure, we can not find a linear boundary through the origin with zero error.

(c) **(5 pts)** We now consider adding a version of Lasso regularization

$$log(Likelihood) - \lambda(|\beta_1| + |\beta_2|)$$

In other words, we regularize the parameters associated with X_1 and X_2, but not β_0. Circle the correct statement below and justify:

First β_1 will become 0, then β_2.

First β_2 will become 0, then β_1.

β_1 and β_2 will become zero simultaneously.

No parameter will be exactly zero, only smaller as λ increases.

Answer:

First β_1 will become 0, then β_2.
X_1 has lower classification ability. Therefore, as λ increases β_1 will be set to 0 first. As λ increases further, β_2 will eventually also become zero.

(d) **(5 pts)** For the regularization problem above, when λ is large, $\beta_1 = 0$ and $\beta_2 = 0$. What is the value of β_0?

Answer:

The dataset has an equal number of cases from each class. When $\beta_1 = 0$ and $\beta_2 = 0$,

$$P\{Y = 1\} = \frac{1}{1 + exp\{-\beta_0\}} = 0.5$$

Therefore, $\beta_0 = 0$