CS6140

Midterm

February 28, 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	10	
2	10	
3	7.5	
4	10	
5	10	
6	5	
7	10	
8	10	
9	10	
10	15	
11	7.5	
12	10	

1. **(10 pts)**

Consider linear regression with one predictor, and no intercept

$$Y_i = \beta X_i + \epsilon_i,\ \epsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Assume that we have a training dataset of n pairs (X_i, Y_i) for $i = 1 \dots n$, and known. Which ones of the following statements correctly represents the maximum likelihood estimation of β? Circle **TRUE** or **FALSE** to each one, and provide an explanation next to it. More than one statement can be true.

$\arg\max_\beta \sum_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \beta X_i)^2\right)$ **TRUE** **FALSE**

$\arg\max_\beta \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \beta X_i)^2\right)$ **TRUE** **FALSE**

$\arg\max_\beta \sum_i \exp\left(-\frac{1}{2\sigma^2}(Y_i - \beta X_i)^2\right)$ **TRUE** **FALSE**

$\arg\max_\beta \prod_i \exp\left(-\frac{1}{2\sigma^2}(Y_i - \beta X_i)^2\right)$ **TRUE** **FALSE**

$\arg\max_\beta \frac{1}{2} \sum_i (Y_i - \beta X_i)^2$ **TRUE** **FALSE**

$\arg\min_\beta \frac{1}{2} \sum_i (Y_i - \beta X_i)^2$ **TRUE** **FALSE**

Answer:

The maximum likelihood function is:

$$\arg\max_\beta L(\beta, \sigma^2, y, x) = \arg\max_\beta \frac{1}{(\sqrt{2\pi\sigma^2})^n} exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - X_i\beta)^2) \tag{1}$$

So the first statement is FALSE because the sum should be inside the exp. The second statement is TRUE because if you do the product in the second statement you will obtain equation (1). The third statement is FALSE the sum should be inside the exp. The fourth statement is equivalent to the second statement, because we are calculating the maximum over β and $\frac{1}{\sqrt{2\pi\sigma^2}}$ does not depend on β and hence we can remove it from the second statement to obtain the fourth statement.

Note that in order to optimize the expression in (1) it suffice to calculate

$$\begin{aligned}
\arg\max_\beta \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - X_i\beta)^2\right) &= \\
\arg\max_\beta \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - X_i\beta)^2\right) &= \\
\arg\max_\beta \left(-\frac{1}{2} \sum_{i=1}^{N} (Y_i - X_i\beta)^2\right) &= \\
\arg\min_\beta \left(\frac{1}{2} \sum_{i=1}^{N} (Y_i - X_i\beta)^2\right)
\end{aligned}$$

This shows that the fifth statement is FALSE and the sixth statement is TRUE.

2. **(10 pts)** Explain the difference between probability and likelihood.

Answer:

Probability is a property of a random variable, given parameters. It describes the chance that a future observation from this random variable will take a particular value (for a discrete random variable), or will belong to an interval (for a continuous random variable). Probability calculations rely on probability distribution function, where the observations are unknown, but parameters are known.

Likelihood is a function of parameters, given previous observations. We use likelihood to find optimal values of the parameters, given the data. As above, likelihood optimizations rely on probability distribution function, however here the observations are known, but parameters are unknown.

3. **(7.5 pts)** Consider linear regression with n observations and p predictors. We can increase the flexibility of the model, by adding higher order terms and statistical interactions. For each of parts (a) through (d) below, indicate whether we would generally expect the performance of regression with higher order terms and interactions to be better or worse than the model with the original predictors. Justify your answer.

   (a) **(2.5 pts)** The sample size n is large, and the number of predictors p is small.
   Answer:
   Likely better: higher order terms may more accurately describe the true relationship and improve the predictive ability. Overfitting is not of an issue if n is very large.

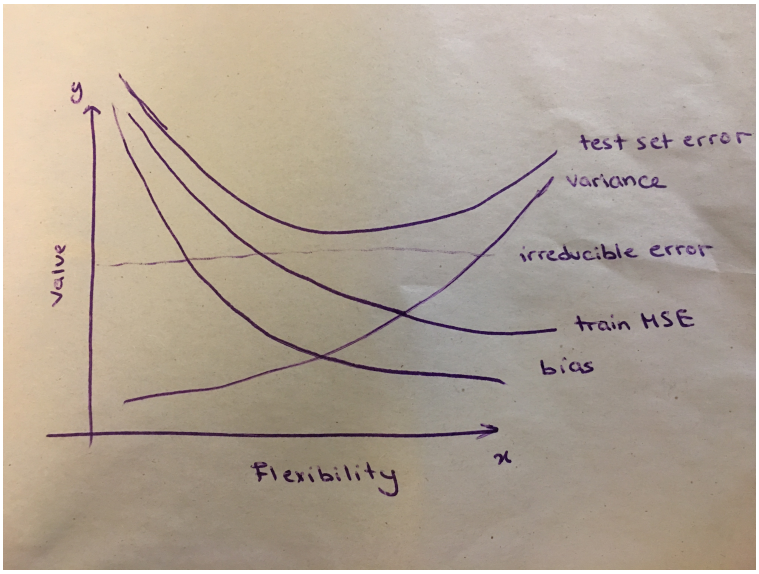
   (b) **(2.5 pts)** The number of predictors p is large, and the number of observations n is small.
   Answer:
   Likely worse: higher order terms will likely due to overfitting, when n is small

   (c) **(2.5 pts)** The variance of the error terms σ^2 is extremely high.

   Answer:
   Likely worse: when the variance of the errors is high and the number of observations is small, a flexible model is more likely to mistake random variation for a systematic signal

4. **(10 pts)** For the same setting as in Problem 2 above, sketch a plot where you overlay 5 curves. The x axis of the plot is the complexity of the linear regression model (i.e., the increasing number of predictors, as more higher order terms are added to the model). For each curve, the y axis of represents (1) bias, (2) variance, (3) training set error, (4) test set error, and (5) irreducible error. Explain the reason for the shapes.

Answer:



Bias: *If a complicated process is represented with a simple linear regression, the regression fails to account for all the systematic effects, and the result is biased. As the flexibility of the model increases, the bias reduces.*

Variance: *A highly flexible model will explain most the points in the data. However, when we collect a new set of observations from the same process, the pattern of the observations will change, and the model will have a different form. In other words, the systematic part of the model will vary a lot between sets of observations. This is referred to as "large variance". When the model has relatively few parameters it is relatively inflexible, and the fit will be similar between replicate sets of observations. In other words, this fit has "small variance".*

Irreducible error: *A constant representing the random variation of the underlying process. It cannot be reduced by statistical modeling, and is the upper bound for accuracy of any model.*

Training set MSE: *In general,* $MSE = Bias^2 + Variance + Irreducible\ error$. *On the training set, we underestimate the variance and the irreducible error. Therefore, training set MSE us related to the bias, and decreases with the model complexity.*

Test set MSE: *On the validation set, we can estimate all the components of the MSE. The result is a U shaped curve. The minimum occurs when both the bias and the variance are minimized (i.e., when the bias line intersects the variance line).*

5. **(10 pts)** The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) **(2.5 pts)** Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

Hint: Eucledian distance between $a = (a_1, a_2, a_3)$ *and* $b = (b_1, b_2, b_3)$ *is* $\sqrt{\sum_{j=1}^{3}(a_j - b_j)^2}$

Answer:

$$\begin{aligned}
Red: \sqrt{(0-0)^2 + (0-3)^2 + (0-0)^2} &= \sqrt{9} = 3 \\
Red: \sqrt{(0-2)^2 + (0-0)^2 + (0-0)^2} &= \sqrt{4} = 2 \\
Red: \sqrt{(0-0)^2 + (0-1)^2 + (0-3)^2} &= \sqrt{10} = 3.16 \\
Green: \sqrt{(0-0)^2 + (0-1)^2 + (0-2)^2} &= \sqrt{5} = 2.236 \\
Green: \sqrt{(0+1)^2 + (0-0)^2 + (0-1)^2} &= \sqrt{2} = 1.41 \\
Red: \sqrt{(0-1)^2 + (0-1)^2 + (0-1)^2} &= \sqrt{3} = 1.732
\end{aligned}$$

(b) **(2.5 pts)** What is your prediction with $K = 1$? Why?

Answer:

$K = 1$

Test set $X1 = X2 = X3 = 0$. *This is close to Green which is at a distance* $\sqrt{2}$. *Therefore the prediction is Green.*

(c) **(2.5 pts)** What is your prediction with $K = 3$? Why?

Answer:

$K = 3$

Test set $X1 = X2 = X3 = 0$. Closest ones are Red (Obs 2), Green (Obs 5) and Red (Obs6). The Red observations are the most common, therefore the prediction is Red.

(d) **(2.5 pts)** Explain the difference between the KNN classifier and the KNN regression.

Answer:

When the labels of the observations are discreet (e.g., Red and Green in this example), we have a classification problem. When the labels are continuous, we have a regression problem.

6. **(5 pts)** For two random variables X_1 and X_2, prove that $P(X_1|X_2)P(X_2) = P(X_2|X_1)P(X_1)$. *Hint: This is a one- or two-line proof.*

Answer:

$$P(X1|X2)P(X2) = \frac{P(X1, X2)}{P(X2)}P(X2) = P(X1, X2)$$
$$P(X2|X1)P(X1) = \frac{P(X1, X2)}{P(X1)}P(X1) = P(X1, X2)$$

Hence, $P(X_1|X_2)P(X_2) = P(X_2|X_1)P(X_1)$.

7. **(10 pts)** Suppose that we wish to predict whether a given stock will issue a dividend this year ('Yes' or 'No') based on the predictor X defines as last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was 10, while the mean for those that didn't was 0. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80 % of companies issued dividends.

   Assuming that X follows a Normal distribution, predict the probability that this year a company will issue a dividend, given that last year its percentage profit was $X = 4$.

   Answer:

$$P\{YES\} = 0.8,\ P\{NO\} = 1 - P\{YES\} = 0.2$$
$$X|Yes \sim \mathcal{N}(10, 6^2),\ X|No \sim \mathcal{N}(0, 6^2)$$

   Using equation 4.12 of JWHT:

$$P\{YES|X=4\} = \frac{P\{YES\} \cdot P\{X=4|YES\}}{P\{YES\} \cdot P\{X=4|YES\} + P\{NO\} \cdot P\{X=4|NO\}}$$
$$= \frac{0.8 \cdot \exp(-(1/72)(4-10)^2)}{0.8 \cdot \exp(-(1/72)(4-10)^2) + 0.2 \cdot \exp(-(1/72)(4-0)^2)} = 0.752$$

   Therefore, the probability that a company will issue a dividend this year given that its percentage return was $X = 4$ last year is 0.752.

8. **(10 pts)** In this problem we consider differences between LDA and QDA.

   (a) **(5 pts)** If the true decision boundary between two classes is linear, do we expect LDA or QDA to perform better on the training set?

   Answer:
   If the Bayes decision boundary is linear, we expect QDA to perform better on the training set because its higher flexiblity may yield a closer fit. On the test set, we expect LDA to perform better than QDA, because QDA could overfit the Bayes decision boundary.

   (b) **(5 pts)** In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

   Answer:
   Improve. QDA is more flexible than LDA and so has higher variance. However, if the training set is very large, the variance of the classifier is not a major concern, but additional flexibility if a plus.

9. **(10 pts)** Consider linear regression, and its parameter estimation. Indicate which of items (i.) through (iv.) are correct, and justify your answer. The lasso, relative to least squares, is:

   (i.) More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   (ii.) More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

   (iii.) Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   (iv.) Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

   Answer:
   (iii) is the correct answer. When the number of predictors is large relative to the sample size, the least square fit maay be negatively affected by uncertainty in parameter estimation. The lasso is a more restrictive parameter estimation approach, and therefore it has less uncertainty. Lasso proceeds by introducing bias, but in return can reduce overfitting and variance in predictions. If the bias due to its added constraints is not too high, lasso will outperform least squares.

10. **(15 pts)** Consider a logistic regression with p predictors, and its parameter estimation with ridge regularization.

(a) **(2.5 pts)** Write the model underlying logistic regression.
Answer:

$$Y \overset{iid}{\sim} Bernoulli(\pi_i) \text{ or } Y \overset{iid}{\sim} Binomial(\pi_i, n_i)$$

$$\pi = P(Y = y|X) = \frac{exp(\beta_0+\beta_1 X_1+...+\beta_p X_p)}{1+exp(\beta_0+\beta_1 X_1+...+\beta_p X_p)} \text{ or } \log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

(b) **(2.5 pts)** Use the notation in (a) to define the criterion that we would like to minimize when working with ridge regularization. *Hint: we will be minimizing the negative log-likelihood, subject to a penalty.*
Answer:

$$\max_{\beta_0,\beta} \left\{ \prod_{i=1}^{N} \left[\pi_i^{y_i} \cdot (1-\pi_i)^{1-y_i}\right] - \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \text{ where} \pi_i \text{ is as in (a)}$$

(c) **(2.5 pts)** True or false: The criterion in (b) has multiple locally optimal solutions.
Answer:
FALSE. It is convex.

(d) **(2.5 pts)** True or false: Parameter estimates optimizing the criterion in (b) are sparse (i.e., they have many zero entries).
Answer:
FALSE. The parameter estimates under the ridge constraint are closer to 0 than the unconstrained parameter estimates, but they are not set to 0. In other words, ridge estimation does not perform variable selection.

(e) **(2.5 pts)** True or false: The log-likelihood corresponding to the optimal criterion in (b) always increases as the regularization parameter increases.
Answer:
FALSE. The larger the regularization parameter, the less flexible is the space of parameter estimates, and the smaller the optimized log-likelihood.

(f) **(2.5 pts)** True or false: In logistic regression (with or without regularization), if we replace the mean response function (or the link function) with the identity function, the parameter estimates and the prediction will be identical to those of linear regression.
Answer:
FALSE. In logistic regression, the maximum likelihood procedure optimizes the Bernouilli (or Binomial) likelihood, and not the residual sum of squares that would be the case in linear regression. Therefore, even with the identity link, the parameter estimates will generally not be the same.

11. **(7.5 pts)** Compare generative classifiers versus discriminative classifiers. For each question, circle the most appropriate answer, and explain the reason. *Hint: you can use logistic regression and LDA as examples.*

    (a) **(2.5 pts)** Which classifier is typically easier to fit?
    Discriminative **Generative**

    Answer:
    Generative. For example, parameters of LDA are obtained separately for each class, in terms of sample means and sample variance-covariances. By contrast, logistic regression requires solving a more complex convex optimization problem.

    (b) **(2.5 pts)** Which classifier is best suited for separately fitting models to each class?
    Discriminative **Generative**

    Answer:
    Generative. As in (a), we estimate the parameters of each class conditional density independently. This is advantageous, e.g. when we add more classes, as we do not need to retrain the model for each class. In contrast, in discriminative models the whole model must be retrained if we add a new class.

    (c) **(2.5 pts)** Which classifier can best handle features fixed by experimental design, or feature transformations?
    Discriminative **Generative**

    Answer:
    Discriminative. A big advantage of discriminative methods is that they do not require probability distribution assumptions for the predictors. Therefore, we can fix predictors at particular values, or use transformations.

12. **(10 pts)** A heuristic for assessing the usefulness of principle component analysis (PSA) on a particular dataset is as follows. Let the empirical variance-covariance matrix of the features Σ have eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$. We can then consider the sample variance of the eigenvalues. Which value of the variance (larger or smaller) indicates the usefulness of PCA on the dataset? Circle the most appropriate answer, and explain the reason.
    Smaller variance **Larger variance**

    Answer:
    Larger. Larger variance means that there is more difference between the values of lambda$_i$, i.e., a larger proportion of total variation is explained by the first few principle components, and last principle components explain a smaller proportion of total variation.