A

Midterm Exam
GU4241/GR5241 Fall 2016

Name _____

UNI _____

Problem 0: UNI (2 points)

Write your name and UNI on the first page of the problem sheet. After the exam, please return the problem sheet to us.

Problem 1: Short questions (2+2+3+3+4+4 points)

Short answers (about one sentence) are sufficient.

(a) (**Yes/No**) Does the optimality of the Bayes classifier depend on the choice of loss function? Explain briefly.

(b) (**Yes/No**) Does the EM algorithm always converge to a global maximizer of the likelihood? Expalin briefly.

(c) Consider a natural cubic spline and a polynomial regression with the same degree of freedom on the same date set. Which is likely to be more stable for extreme values of the predictor?

(d) Which one of the following algorithms does not rely on the normality assumption: K-means, LDA, ordinary least squares, principle components analysis. Explain briefly.

(e) Describe sampling from a finite mixture model $\pi(x) = \sum_{k=1}^{K} c_k p(x|\theta_k)$ as a two-step procedure.

(f) List two regression methods for which it is possible to compute the leave one out cross validation (LOOCV) error analytically without performing n fits.

Solution:

(a) **Yes.** Bayes classifier is optimal under 0-1 loss.

(b) **No.** EM algorithm always converges. But it may converge to a local maxima.

(c) The natural cubic spline. Natural cubic spline requires the fit be linear beyond the boundaries.

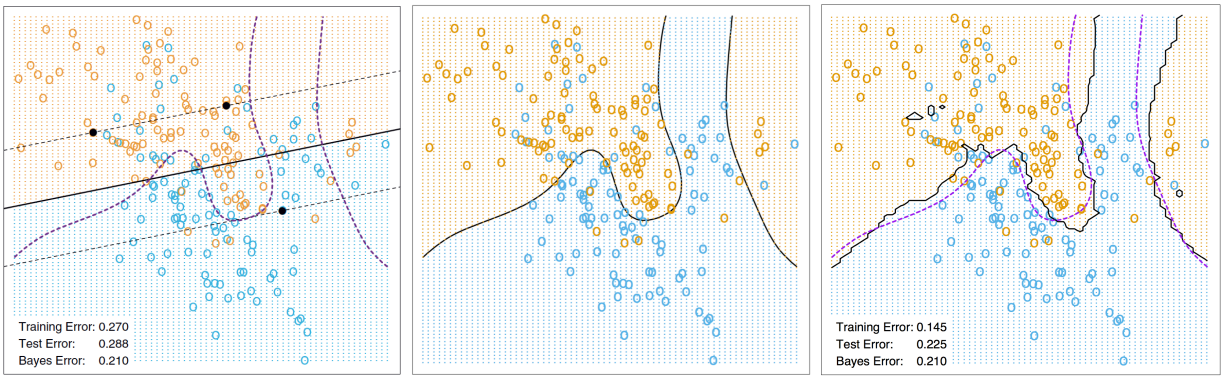
(d) PCA. K-means clustering is a special case of the EM algorithm for mixture models. When the model is finite mixture of Gaussian distributions with the identity as the covariance matrix, EM algorithm is equivalent to k-means. LDA assumes the data from each class follow a Gaussian distribution with the same covariance matrix. Ordinary least squares assumes the errors are i.i.d. normal.

(e)
- Choose a mixture component at random. Each component k is selected with probability c_k.
- Sample x_i from $p(x|\theta_k)$.

(f) Ordinary least squres, ridge regression, cubic splines, natural cubic splines, smoothing splines.

Problem 2: Decision boundaries (10 points)

The following pictures, which we have all seen in class, show the output of several different classifiers. Recall that the thick line is the decision boundary determined by the classifier; you can ignore the dashed lines.



(a) (b) (c)

For each of the three pictures:

- Name at least one classifier which could have produced this solution. Explain why.
- Name at least one classifier which could not have produced the solution. Explain why not.

Solution:

	(a)	(b)	(c)
could be generated by reason	logistic regression or LDA linear boundary, class overlap	Bayes classifier smooth, non-linear boundary	K-nearest-neighbor classifier non-linear boundary
could not be generated by	K-nearest-neighbor classifier Trees or RF (smooth slope)	any linear classifier Trees or RF (smooth)	any linear classifier

Problem 3: K-means clustering (10 points)

Perform K-means clustering on the following 2-dimensional observations with $K = 3$ and initial labels $(1, 1, 2, 3, 3, 2)$. Use the *Manhattan* distance between a pair of points: $d(A, B) = |X_{1A} - X_{1B}| + |X_{2A} - X_{2B}|$, instead of the Euclidean distance. With this distance, the centroid of a cluster is obtained by taking the median of the samples in each dimension. Show your results after each iteration.

Obs.	X_1	X_2
A	1	4
B	1	3
C	3	4
D	5	2
E	3	2
F	3	0

Solution:

The centroids of the three clusters are $(1, 3.5)$, $(3, 2)$ and $(4, 2)$. The distances between the observations and these centroids is shown in the following table:

dist.	C_1	C_2	C_3
A	0.5	4	5
B	0.5	3	4
C	2.5	2	3
D	5.5	2	1
E	3.5	0	1
F	5.5	2	3

So the new labels after the first iteration are $(1, 1, 2, 3, 2, 2)$, and the centroids of the new clusters are $(1, 3.5)$, $(3, 2)$ and $(5, 2)$. Since only one centroid changed, we just need to compute the distance between the third centroid and the data points.

dist.	C_1	C_2	C_3
A	0.5	4	6
B	0.5	3	5
C	2.5	2	4
D	5.5	2	0
E	3.5	0	2
F	5.5	2	4

The clusters remain the same, so the result is $\{A, B\}$, $\{C, E, F\}$ and $\{D\}$.

Problem 4: Logistic Regression (10 points)

Suppose we have a dataset with N observations, and each observation consists of three values:

- y: binary variable that is 1 if a student passed and 0 if a student failed the exam
- x_1: the number of hours spent studying for the exam
- x_2: a binary variable indicating whether or not the student passed the previous exam.

Suppose upon fitting a logistic regression of y on x_1, x_2, and an intercept, the estimates for $\beta = (\beta_0, \beta_1, \beta_2)$ are

$$\begin{aligned} \hat{\beta}_0 &= -1.2 \\ \hat{\beta}_1 &= 0.3 \\ \hat{\beta}_2 &= 1.2 \end{aligned}$$

Now suppose instead of using the number of hours spent studying, we used the number of minutes spent on the exam preparation. Can you identify what the new β_0, β_1, and β_2 would be? Explain your answer.

Solution:

Logistic regression is fit by maximizing the likelihood function, which can be written in the following way:

$$\hat{\beta} = \arg\min_{\beta} \prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}} \prod_{i:y_i=0} \frac{1}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}}.$$

This only depends on the parameters through a linear function. After we change the unit, we want to solve the following optimization problem:

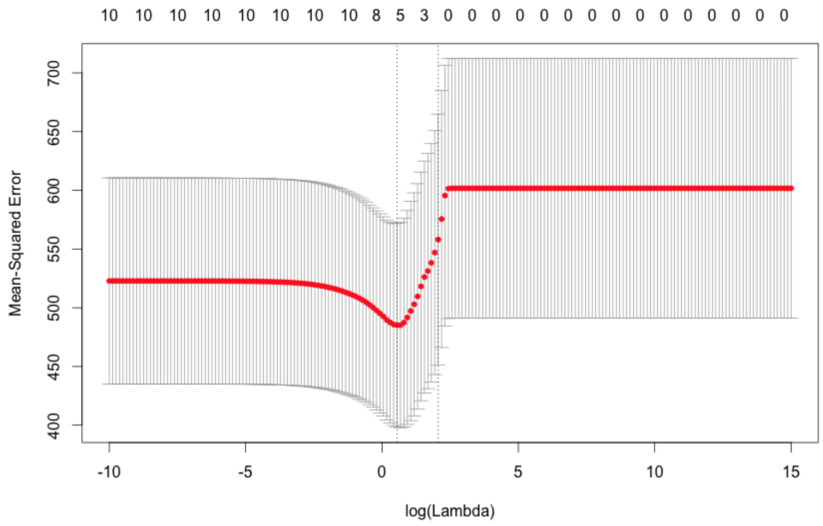
$$\hat{\beta}^{\text{new}} = \arg\min_{\beta} \prod_{i:y_i=1} \frac{e^{\beta_0+\beta_1 60 x_{i1}+\beta_2 x_{i2}}}{1+e^{\beta_0+\beta_1 60 x_{i1}+\beta_2 x_{i2}}} \prod_{i:y_i=0} \frac{1}{1+e^{\beta_0+\beta_1 60 x_{i1}+\beta_2 x_{i2}}}.$$

The maximizers $\hat{\beta}$ and $\hat{\beta}^{\text{new}}$ are related by

$$\begin{aligned} \hat{\beta}_0^{\text{new}} &= \hat{\beta}_0 = -1.2, \\ \hat{\beta}_1^{\text{new}} &= \frac{1}{60}\hat{\beta}_1 = 0.005, \\ \hat{\beta}_2^{\text{new}} &= \hat{\beta}_2 = 1.2. \end{aligned}$$

Problem 5: Tuning parameter of lasso

The plot below displays the cross-validation errors of lasso regression computed on a range of tuning parameters λ. There are $p = 10$ parameters in the model, and the training set and test set has $n = 100$ observations each.



In terms of the cross validation error, what properties do the two models marked with dashed vertical lines in the plot have? Would you expect an appropriate shrinkage method to give a smaller test MSE than the test MSE of multiple linear regression for this data?

Solution:

The two dashed lines mark the model with the smallest cross validation error and the model chosen with the one standard error rule; that is, the simplest model whose cross validation error is within one standard error of the minimum cross validation error. The lasso regression when $\lambda \to 0$ corresponds to the linear regression with $(10 + 1)$ parameters including the intercept and the lasso regression when $\lambda \to \infty$ correponds to the linear regression with intercept only. We observe that the lasso regression with a certain tuning parameter has the smallest cross validation error. Since the mean-squared errors from cross validation are estimates of the test MSE, we expect that the lasso regression using the best λ from cross-validation will give a smaller test MSE than the test MSE of multiple linear regression.

Problem 6: Variable selection (10 points)

Your colleague fitted a multivariate linear regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$ and found all but three p-values are significant in the t-test. He decides to drop those three variables and keep all the remaining predictors. Do you think this is a good idea? Briefly explain your answer. If you think this method is inappropriate, could you suggest an alternative?

Solution:

The colleague's method is not reasonable because the t-test only tests the marginal effect of each predictor. If there is collinearity, it is possible that after removing one predictor, formerly insignificant predictors become significant. If p is large, choosing significant predictors at a fixed significance level could also lead to a large number of false positives. A better approach would be to apply forward or backward stepwise selection, then use cross-validation to select the optimal model.

Problem 7: Step function regression (10 points)

We fit a step function regression on a dataset with a single predictor X. 3 knots c_1, c_2, c_3 in the range of X are selected. We construct 4 variables

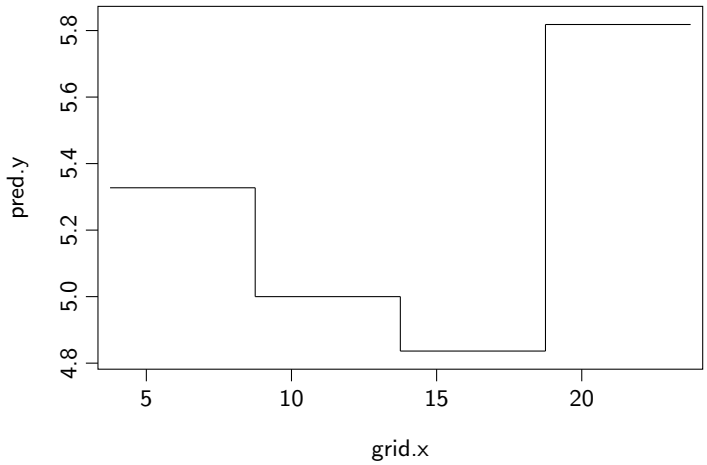
$$C_0(X) = \mathbb{1}_{\{X<c_1\}}, \quad C_1(X) = \mathbb{1}_{\{c_1\leq X<c_2\}}, \quad C_2(X) = \mathbb{1}_{\{c_2\leq X<c_3\}}, \quad C_3(X) = \mathbb{1}_{\{c_3\leq X\}},$$

where $\mathbb{1}_{\{.\}}$ is an indicator function. For example, C_0 takes value 1 when $X < c_1$ is true; and is equal to 0 otherwise. Note that the linear model using $C_0(X), C_1(X), C_2(X), C_3(X)$ as predictors is:

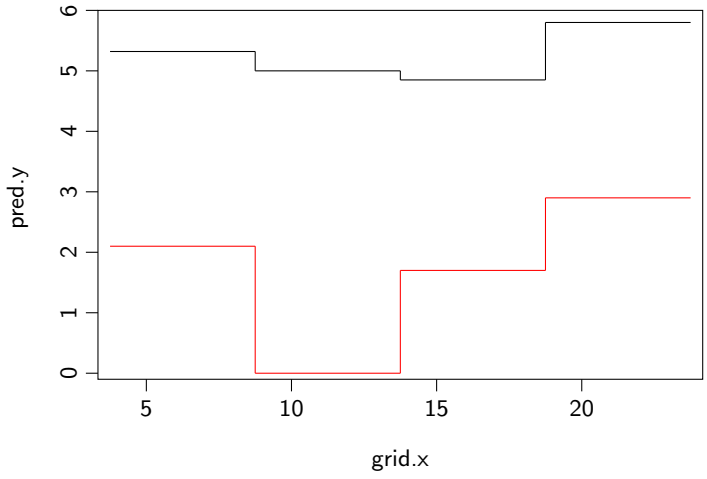
$$Y = \beta_0 C_0(X) + \beta_1 C_1(X) + \beta_2 C_2(X) + \beta_3 C_3(X) + \epsilon, \quad \epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).$$

The plot below displayes the fitted step function on a dataset.

Now we fit a lasso regression using the same predictors. Using the tuning parameter $\lambda = 0.17$, we get 3 nonzero β_is out of 4 and $\hat{\beta}_1 = 0$. Sketch the lasso-fitted step function on the plot below. Briefly explain how you derive this lasso solution.



Solution:
If we fit a lasso regression using the same predictors, the estimates will be shriked to zero. We already know that $\hat{\beta}_1 = 0$, which implies that the fitted value is zero when $c_1 \leq x < c_2$. The sketh of the lasso-fitted step function is shown by the red curve in the plot below. The black curve represents the fit without the L_1 penalty.



Problem 8: Principle components regression (10 points)

The singular value decomposition (SVD) of the input matrix $\mathbf{X}$ provides us some insights into the nature of the regression methods. We assume that the input matrix $\mathbf{X}$ is centered and appropriately scaled. The SVD of the $N \times p$ matrix $\mathbf{X}$ has the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T.$$

We assume all singular values d_j are positive. Using the singular value decomposition we can write the least squares fitted vector as

$$\begin{aligned} \mathbf{X}\hat{\beta}^{\text{ls}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^{p} \mathbf{u}_j\mathbf{u}_j^T\mathbf{y}, \end{aligned}$$

where $\mathbf{u}_j$ is the jth colume of $\mathbf{U}$. This expression means we can obtain the fitted value directly if we use the orthonormal basis $\mathbf{u}_1, \cdots, \mathbf{u}_p$. $\mathbf{u}_1^T\mathbf{y}, \cdots, \mathbf{u}_p^T\mathbf{y}$ are the coordinates of $\mathbf{y}$ with respect to the new basis.

(a) We know that the ridge solutions are

$$\mathbf{X}\hat{\beta}^{\text{ridge}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

Similar to the least squares, can you rewrite the expression using the SVD of $\mathbf{X}$?

(b) The idea of principle components regression is that we replace the original inputs X_j by a small set of linear combinations of X_j based on the principle components directions v_m (v_m is the mth column of $\mathbf{V}$). More precisely, we derive the input columns $\mathbf{z}_m = \mathbf{X}v_m$, and then regress $\mathbf{y}$ on $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_M$ for some $M \leq p$. If we write our estimate as

$$\hat{\mathbf{y}}^{\text{pcr}} = \sum_{m=1}^{M} \theta_m \mathbf{z}_m,$$

what are the estimated $\hat{\theta}_m$s which miminze the residule sum of squares?

(c) Could you list some pros and cons of principle components regression?

Solution:

(a)

$$\begin{aligned} \mathbf{X}\hat{\beta}^{\text{ridge}} &= (\mathbf{U}\mathbf{D}\mathbf{V}^T)(\mathbf{V}\mathbf{D}^2\mathbf{V}^T + \lambda\mathbf{I})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j^T\mathbf{y}. \end{aligned}$$

(b) Note that $\mathbf{z}_m$ are orthogonal. Therefore

$$\hat{\theta}_m = \frac{\langle \mathbf{y}, \mathbf{z}_m \rangle}{\langle \mathbf{z}_m, \mathbf{z}_m \rangle}.$$

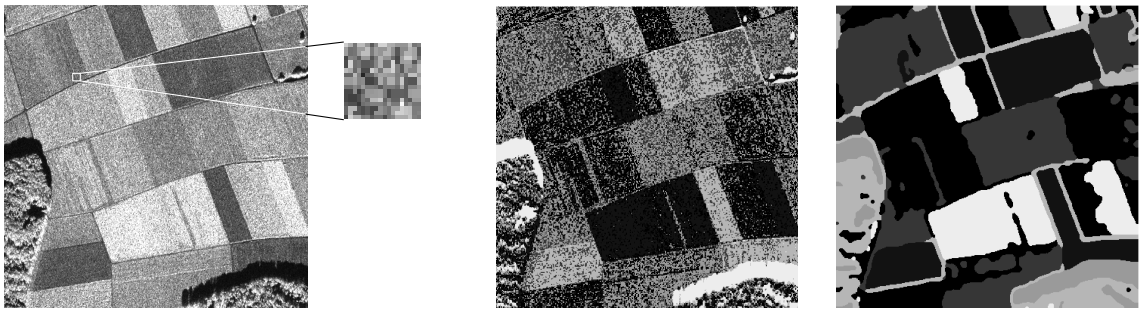
(c) Pros:

- PCR can provide stable estimate when $\mathbf{X}^T\mathbf{X}$ is almost singular.
- PCR can reduce the dimension of the problem.

Cons: it is difficult to interpret the model.

Problem 9: Histogram clustering (5+5 points)

In homework 1, you implemented a histogram clustering algorithm for image segmentation. Recall that the histograms you used had been extracted by placing a small window at regularly spaced points in the image and computing a histogram from the points inside the window (figure (i) below):



(i) (ii) (iii)

The two segmentation solutions (ii) and (iii) have been obtained by the same algorithm on the same image, with histograms extracted at the same locations. The only difference between the two data sets is the size of the histogram windows: One uses histogram over small windows (3×3 pixels), one is based on large windows (19×19 pixels). The centers of neighboring windows are 4 pixels apart.

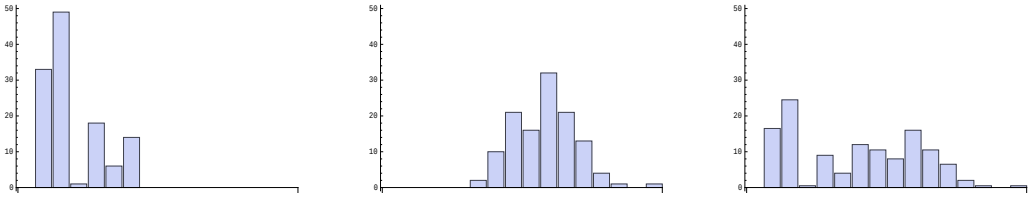
(a) Compare the segmentation solutions (ii) and (iii). Can you tell which one has been generated using small histogram windows, and which using large ones? Please explain your answer.

(b) In solution (iii), you can see that the borders between some neighboring segments become a segment of their own. Can you explain why that happens?

Solution:

(a) Small histogram windows: (ii) Large histogram windows: (iii)
The large histograms windows have a large overlap (the overlap with any neighboring window is almost half the window size). The distributions represented by the histograms are therefore more similar (they share almost half their data points), which results in smoother segmentation results.

(b) At the boundary, the large window overlaps two segments, and the histogram extracted from this window is hence a mixture of the distributions of the segments. For example:



Segment 1 Segment 2 Mixed

If the segment distributions differ significantly, the mixed distribution differs significantly from both, and becomes a segment in its own right.