A

Midterm Exam

GU4241/GR5241 Fall 2016

Name

____

UNI

____

Problem 0: UNI (2 points)

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

Problem 1: Short questions:

(a) False

(b) False, the optimality of the Bayes classifier is under 0-1 loss.

(c) We will choose EM algorithm, because the variances are different. K-means is equivalent to the EM algorithm if we use Gaussian mixture model and assume the covarian matrix is I.

(d) In addition to reducing the variance of the regression estimate through shrinkage, the Lasso performs variable selection by setting certain coefficient to 0.

(e) An outlier is a point for which y_i is far from the value predicted by the model. We can look at the studentized residual to check if a point is an outlier. In order to know if (x_i, y_i) has high leverage, we need to compute its leverage statistic h_i.

(f) If there were no continuity constraints, we would have 4 parameters for each cubic piece, for a total of $4(K+1)$ paremeters. However, we have 3 constraints for each knot, and each one deliminates one degree of freedom. This leaves us with $4K + 4 - 3K = K + 4$ parameters.

Problem 2: Hierarchical clustering:

We compute the distance matrix between the samples:

dist.	A	B	C	D	E	F
A		1	2	6	4	6
B			3	5	3	5
C				4	2	4
D					2	4
E						2
F						

The smallest distance is that between A and B, so we form a new cluster {A, B}. Then, we recompute the distance matrix between our current set of clusters:

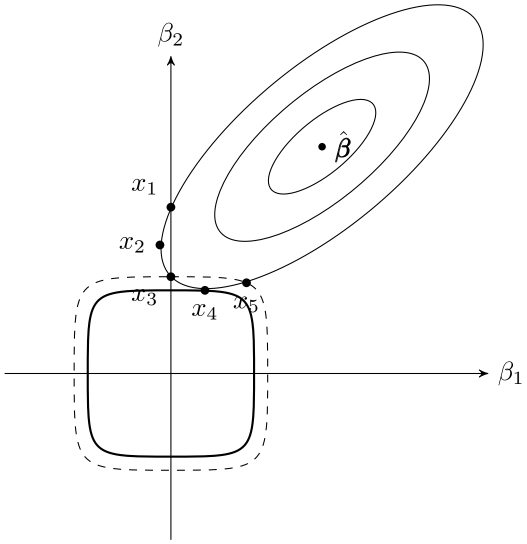
dist.	{A, B}	C	D	E	F
{A,B}		2	5	3	5
C			4	2	4
D				2	4
E					2
F					

At this second step, we link all the clusters, because every cluster is within a distance 2 to another cluster. We obtain the single cluster {A, B, C, D, E, F}.

Problem 3: ℓ_q-regression:

(a) $q = 0.5$ encourages sparse solutions, wheras $q = 4$ does not.
Explanation: If $q = 0.5$, then for any ellipse iso-line of the square-loss which interesects an axis, the minimal $\ell_{0.5}$-distance to the origin (and hence to the smallest penalty term) is achieved by a point on the axis. Hence, the entry for the respective other axis is zero. In contrast, $q = 4$ encourages entries β_i of roughly even size.

(b) $q = 0.5$: The cost-optimal point is x_3, since it is located on the β_2-axis as explained above.
$q = 4$: The cost-optimal point is x_4, since we can shrink the penalty iso-line in the figure further until it intersects the ellipse only at x_4.



Problem 4: Logistic regression:

Notice that y is labeled 'slow' if and only if $x < 4$. That is, the two classes can be perfectly seperated. We know that in this case, logistic regression produces unstable results. In particular, if we choose $\beta_0 = 0$ and let $\beta_1 \to \infty$, the probability of each observation approaches 1. Therefore, our objective can only get better as we increase β_1, and there is no maximizer of the likelihood.

Problem 5: Cross validation:

This estimate is slightly biased downward. The reason is that we have used all the training data to select the optimal, and then used the same data to estimate the test error with $\lambda = 0.5$. The model with the optimal λ is already fit to all the data, so the left-out folds are not entirely independent of the model. A better estimate of the test MSE could be produced by first splitting the data into training and test sets, selecting the optimal λ by 10-fold cross validation on the training set, and calculating the test MSE on the test set.

Problem 6: Variable selection:

This problem is in the realm of high-dimensional statistics, since p is significantly larger than n. Even though some antibodies may be predictive of the disease, adding predictors which are not correlated with the response may hurt the performance of the method. This is because the number of patients is relatively small and their data are used to perform variable selection, which is more difficult if the significant variables are obscured by many insignificant predictors.

Problem 7: Polynomial regression:

(a) The solution for a point x is

$$f(x) = \beta_0 + \beta_1 x.$$

For the two specific points in the problem that is

$$f(x) = -1 + 4 \cdot 0.8 = 2.2,$$

and

$$f(x') = -1 + 4 \cdot 1.0 = 3.$$

(b)

$$\hat{\beta}_{\text{OLS}} = (X^T X)^{-1} X^T y.$$

β is a $d+1$ dimensional vector.

(c) The solution for a point x is

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2.$$

For the two specific points in the problem that is

$$f(x) = -1 + 4 \cdot 0.8 + 1 \cdot 0.64 = 2.84,$$

and

$$f(x') = -1 + 4 \cdot 1.0 + 1 \cdot 1.0 = 4.$$

(d) Let Φ be the augumented data matrix with $2d+1$ dimensions, corresponding to the intercept term, d linear terms, and d simple quadratic terms. Then, we perform least squares on this new feature matrix:

$$\hat{\beta}^{\text{quad}}_{\text{OLS}} = (\Phi^T \Phi)^{-1} \Phi^T y.$$

Also, $\beta^{\text{quad}} \in \mathbb{R}^{2d+1}$.

(e)

$$\hat{\beta}^{\text{quad}}_{\text{Ridge}} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T y.$$

Again, $\beta^{\text{quad}} \in \mathbb{R}^{2d+1}$.

Problem 8: Model evaluation for clustering:

We can randomly split the data into two sets $\mathcal{X}'$ and $\mathcal{X}''$ of equal size. First we apply hierarchical clustering to each data set seperately, and obtain results π' and π'' (both π' and π'' are mappings from $\{1, 2, \ldots, n\}$ to $\{1, 2\}$). Assume that for each $x_i \in \mathcal{X}''$, the assignments under π'' is m_i''.

Then we will use π' to "make predictions" on $\mathcal{X}''$. Given the two clusters produced by π', assign each $x_i \in \mathcal{X}''$ to the closest cluster under the same metric. Denote the assignment as m_i'.

We define a statistic as the following:

$$\psi(\mathcal{X}) = \min_{\sigma} \sum_{i=1}^{n} \mathbb{I}\{m_i' \neq \sigma(m_i'')\}$$

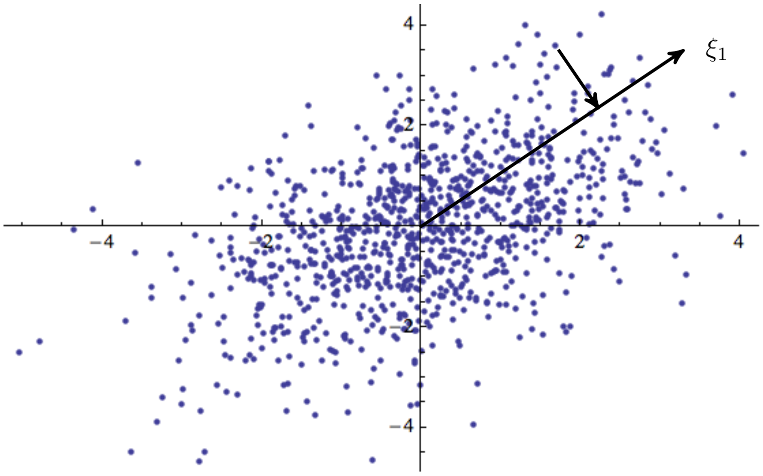
where the minimum is over the permutations σ which permute $\{1, 2\}$.

In order to calculate the p-value, we randomly select b different sets of 1000 genes, repeat the previous steps and caculate the statistic ψ. The p-values can be otained based on the quantitles of ψ.

Problem 9: PCA:

Points: (a) 2+2 points; (b.i) 2 points; (b.ii) 4 points.

(a)



The opposite orientation of ξ_1 is also correct, of course.

(b) (i) There are $d := 10304$ principal components.

(ii) A representation of the image x in terms of the principal components can be computed by projecting x onto each principal component ξ_j to obtain a coefficient c_j:

$$c_j := \langle x, \xi_j \rangle$$

The image can then be represented by an expansion in the basis $\{\xi_1, \ldots, \xi_d\}$:

$$x = \overline{x} + \sum_{j=1}^{d} c_j \xi_j$$

To reconstruct x approximately from the first 48 components, we truncate the expansion at $j = 48$:

$$\hat{x} = \overline{x} + \sum_{j=1}^{48} c_j \xi_j$$