Midterm

SDS383C

Fall 2016

You have 75 minutes. The exam is out of 30 points.

Good Luck!

Name: ____

UTeid: ____

Part I: Short questions (10 points)

You should answer the following with at most two sentences; you can use a picture if you want. If your answer is true, give a brief explanation. If you answer false, provide explanation or give a counter-example.

(1 pts) The maximum likelihood estimate of the model parameter α_1 can be learned using linear regression for the model $y_i = \alpha_1 e^{X_{i1}+2X_{i2}} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ are iid noise. Yes. Because y is a linear function of α_1. You can just use $e^{X_{i1}+2X_{i2}}$ as a feature.

(1 pts) The maximum likelihood estimates of the model parameters (α_1, α_2) can be learned using linear regression for the model $y_i = X_{i1}^{\alpha_1} 2^{\alpha_2} + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ are iid noise. No. Because y is a nonlinear function of α_1. Even if you take a log, you are lost, because the errors wont be normal.

(1 pts) The maximum likelihood estimates of the model parameters (α_1, α_2) can be learned using linear regression for the model $y_i = \log(X_{i1}^{\alpha_1} 2^{\alpha_2}) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ are iid noise. Yes, again, y is a linear function of α.

(4 pts) Consider a linear regression problem with two parameters β_0 and β_1. We have n datapoints $(x_1; y_1), \ldots, (x_n; y_n)$. x_i is a scalar. $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed as:

$$(\hat{\beta}_0, \hat{\beta}_1) \leftarrow \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Check which of the following statements are true. Show your work. More than one may be true.

(a) $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) y_i = 0$

(b) $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(\hat{\beta}_0 x_i^2 - \hat{\beta}_1 \bar{y}) = 0$

(c) $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(\hat{\beta}_0 x_i - \hat{\beta}_1 \bar{y}) = 0$

(d) $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i - \bar{x}) = 0$

Consider the derivatives of the objective w.r.t β_0 and β_1. This gives:

$$\sum_i (y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\sum_i (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

So any linear combination of the above two will be zero. Number (c) and number (d) are both linear combinations of these.

(3 pts) You are conducting a Kaggle challenge where several groups are running their algorithm on the same dataset. Which results will you accept? Give a 1/2 line explanation.

accept/disqualify "Our algorithm is the super awesome coolest one. It has the best training error." Disqualify. Training error does not measure predictive power. One may overfit.

accept/disqualify "Our algorithm is the super awesome coolest one. It has the best test error among all other methods. The tuning parameter we chose for our Lasso algorithm was $\lambda = 1.676299211788$". Disqualify. Seems like λ was very specific, so they probably picked it by maximizing test error. Sounds fishy. Anyone who answered accept, and "I would like to know if they did CV " got full points.

accept/disqualify "Our algorithm is the super awesome coolest one. It has the best test error among all other methods. We report the results for the best value of λ." Disqualify. Its not clear how λ was picked. Did they do cross validation or picked λ that maximizes test error.

Part II: Long questions

1. **Estimation and robustness** (6 points)

   Consider a dataset where with probability $1-\alpha$, a datapoint comes from $Uniform([0, \theta])$ and with probability α it can come from any arbitrary distribution. All n datapoints are i.i.d.

   (a) (1 pt) You know $\alpha = 0$. So there is no contamination in your dataset. What is the Maximum Likelihood Estimate of θ? Lets call this $\hat{\theta}$. $\max(x_1, \ldots, x_n)$

   (b) (2 pts) Calculate the sensitivity curve for $\hat{\theta}$. Recall that for an uncontaminated dataset $x_1, \ldots, x_{n-1}$, and an outlier point x, the Sensitivity curve computes

   $$SC(x) = n\left(\hat{\theta}(x_1, \ldots, x_{n-1}, x) - \hat{\theta}(x_1, \ldots, x_{n-1})\right).$$

   $$SC(x) = \begin{cases} 0 & \text{If } x \geq \max(x_1, \ldots, x_{n-1}) \\ x - \max(x_1, \ldots, x_{n-1}) & \text{o.w.} \end{cases}$$

   (c) (1 pts) Based on the value of the Sensitivity curve, do you think $\hat{\theta}$ is appropriate if there were outliers, i.e. $\alpha > 0$? Explain your answer. $SC(x)$ depends on x, so for a large x it can be unbounded. Hence its not appropriate for outliers.

   (d) (2 pts) You happen to know that $\alpha = .01$. Now construct a robust variant of the estimator of θ. Explain your answer. Just take the $\lfloor n \times .99 \rfloor^{th}$ order statistic. Those who replied trim data by taking .05% off both ends lost some points.

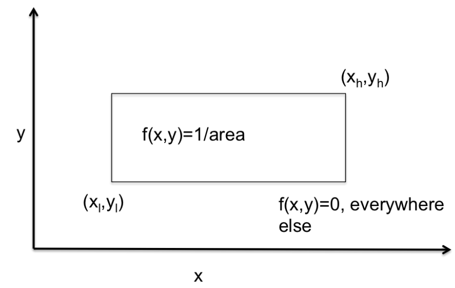


Figure 1: Rectangular bivariate uniform distribution

2. **Classification** (14 points)

   We will define the density of a bivariate uniform distribution (Figure 2) over a rectangle $(X, Y) \sim R(x_\ell, y_\ell, x_h, y_h)$ as:

$$f_{X,Y}(x, y) = \frac{1(x_\ell \leq x \leq x_h, y_\ell \leq y \leq y_h)}{(x_h - x_\ell)(y_h - y_\ell)}.$$

   Recall that the marginal pdf of X and Y are also uniform. For concreteness, if $(X, Y) \sim R(0.1, 0, .3, 1)$, $f_{X,Y}(x, y) = 5$ for $x = .2, y = .5$.

   (a) (4 pts) Assume that we have n data points where each data point is an iid draw from $R(x_\ell, y_\ell, x_h, y_h)$. What are the Maximum Likelihood Estimates of x_ℓ, y_ℓ, x_h, y_h? (No need to show derivation.)

$$\hat{x}_h = \max x_1, \ldots, x_n$$

$$\hat{y}_h = \max y_1, \ldots, y_n$$

$$\hat{x}_\ell = \min x_1, \ldots, x_n$$

$$\hat{y}_\ell = \min y_1, \ldots, y_n$$

   Now consider the following example with 10 datapoints.

x	y	Class
0	2	1
1	0	1
1	5	1
2	3	1
8	4	1
6	6	2
7	4	2
5	7	2
6	3	2
0	6	2

Table 1: Table of data points

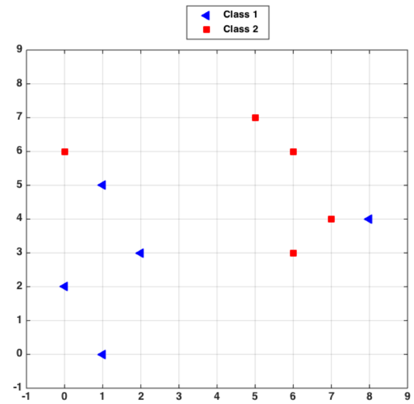
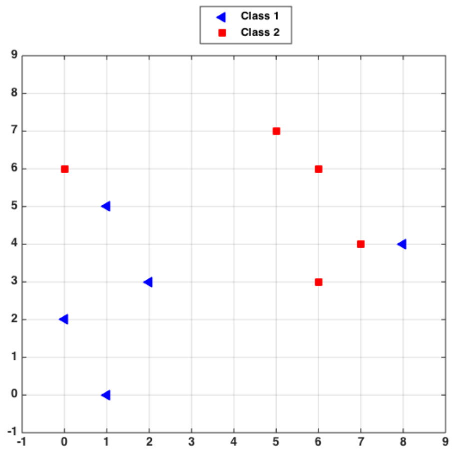


Figure 2: Plot of data points

You believe that datapoints in each class come from a bivariate uniform distribution. Using the data in Table 1 (plot in Figure 2) to estimate parameters of the rectangular bivariate uniform distribution for each class, answer the following questions.

(a) (3.5 pts) Use Bayes rule to classify the point $(0, 1)$. Recall that Bayes rule assigns a point (x, y) to the class k such that $k = \arg\max_{i\in\{1,2\}} P(Class = i|X = x, Y = y)$. $P(X = x, Y = y|class = i) = 1/40$ for $i = 1$ and $1/28$ for $i = 2$. But $P(X = 0, Y = 1|class = 2) = 0$ since it does fall inside the rectangle. Hence obviously it will be assigned to class 1.

(b) (3.5 pts) Use Bayes rule to classify the point $(2, 4)$. Since the class priors are the same, i.e. .5, for a point that belongs to both rectangles, $P(X = x, Y = y|Class = i)$ is maximized for the smaller rectangle. Hence assigned to class 2.



(c) (Outliers)

i. (3 pt) Do you think the data has outliers? If yes, which datapoints do you think are outliers? Explain your answer. $(0, 6)$ and $(8, 4)$ are outliers. Because if they weren't then the points in each rectangle will be a lot more spread out. Many of you have said that the points are far away from the center of the rest etc.