# Midterm

## SDS383C

*Fall 2015*

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

## Good Luck!

**Name:** _____

**UTeid:** _____

# Part I: Short questions (14 points)

1. (4 points) Generate $(Z_1, X_1), \ldots, (Z_n, X_n)$ as follows:

$$Z_i = \begin{cases} \text{“blue” with probability } \theta \\ \text{“red” with probability } 1-\theta \end{cases} \qquad X_i \sim \begin{cases} Poisson(1) & \text{If } Z_i = \text{“blue”} \\ Poisson(4) & \text{If } Z_i = \text{“red”} \end{cases}$$

You are only given the $X_i$'s. Find a consistent estimator of $\theta$ that avoids using EM. Argue why your estimator is consistent.

**Solution**:

Calculate $E[X_i] = E[E[X_i|Z_i]] = \theta 1 + (1-\theta)4 = 4 - 3\theta$. By law of large numbers $\sum_i X_i/n \xrightarrow{P} E[X_i] = E[X]$. So set $\hat{\theta} = (4 - \bar{X})/3$

2. (3 points) Suppose we run a ridge regression with parameter $\lambda$ on $p$ variables $X_1, \ldots, X_p$. The coefficient I estimate for $X_1$ ($\hat{\beta}_{ridge}(1)$) is $a$. Now $m-1$ additional copies of variable $X_1$, i.e. $X_1^* = X_2^* = \cdots = X_{m-1}^* = X_1$ are included and the ridge regression is refit. How are the new coefficients of the identical copies related to $a$ ? Prove your answer. **Solution**: Since the $X^*$'s are identical copies of $X_i$'s and everything else are held fixed, in the alternate problem we are looking for:

$$\arg \min_{\beta_1, \ldots, \beta_m} \sum_i \beta_i^2$$
$$\text{s.t.} \sum_i \beta_i = a.$$

This is minimized when $\beta_1 = \cdots = \beta_m = a/m$,

3. (1 point) In n-fold cross-validation each data point belongs to exactly one test fold, so the test folds are independent. Are the error estimates of the separate folds also independent? So, given that the data in test folds $i$ and $j$ are independent, are $e_i$ and $e_j$ , the error estimates on test folds $i$ and $j$, also independent? Explain briefly. **Solution**: No. The models are dependent since they use the same data to train.

4. (1 points) (True/False) MAP estimates are more prone to overfitting than MLE. Explain. **Solution**: No. MAP estimates smoothes the MLE estimates and in fact prevent overfitting.

5. (1.5 points) Write down 3 examples of estimators you have seen in the class which reduce the overall MSE at the expense of introducing a little bias. **Solution**: James Stein, Ridge, Lasso.

6. (1+2.5) Your evil twin gives you a dataset $Y_1, \ldots, Y_n$. He tells you the points are drawn i.i.d from a normal distribution, but he will not divulge the mean $\mu$ and variance $\sigma^2$.

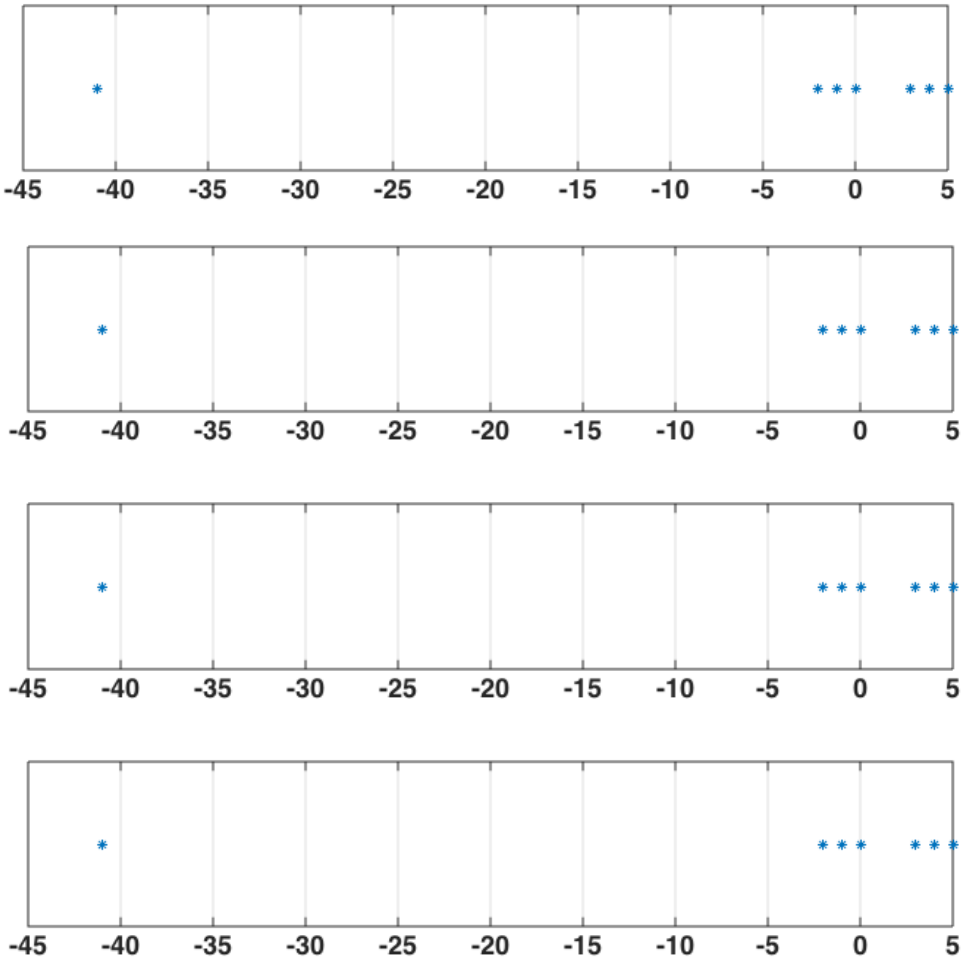
   (a) Write down a consistent estimator of $\sigma$. (No explanation needed). **Solution**: Sample variance (unbiased version)

   (b) We will call your answer from the last part $T_n$. Your evil twin asks you to estimate the variance of $T_n$. Name a procedure you have learned in class to do that. Briefly write down the steps to do it. *Hint: your twin is not evil enough to want you to write long equations.* **Solution**: Parametric bootstrap

# Part II: Long questions

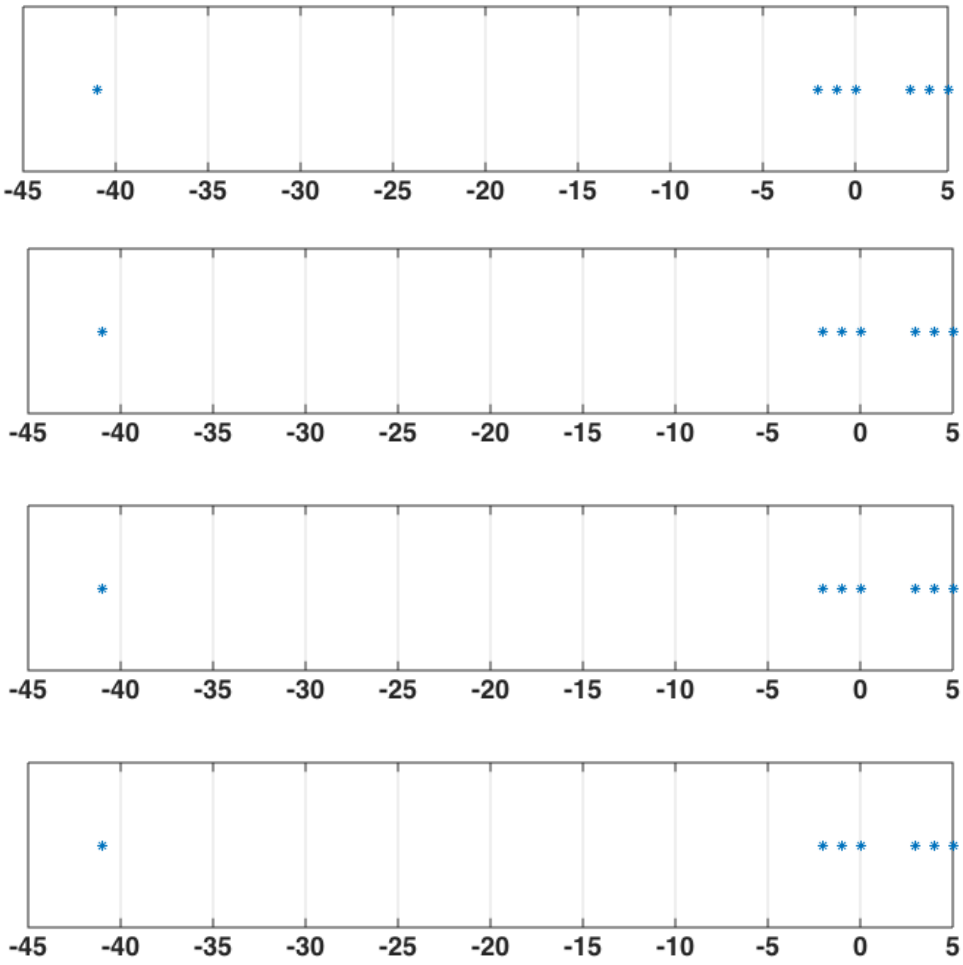
1. **CLUSTERING (8 points)**

   (a) (3 points) You have 7 datapoints $-41, -2, -1, 0, 3, 4, 5$. Run k-means with $k = 2$ and initial centers at $-1$ and 4. Use the following figures to show the successive clusters and cluster centers you get. **Solution**: All of you got this right. So, I will skip.



(b) (1 points) You now have reason to believe that your data has outliers. How will you change your k-means algorithm to be robust towards outliers? Throwing away data-points is not an option. **Solution**: Use median of the points instead of mean!

(c) (3 points) Now show the steps of your new algorithm over the same dataset with the same initialization.



(d) (1 point) You are doing a summer project with Prof. Sarkar. She has a set of microarray data with a million genes, and 98 patients, with continuous outcome (survival time) for each patient. There are no missing data. Your fast computer has broken and would not be ready until the end of summer. Since you are stuck with a slow computer, Professor Sarkar asks you to first select the 1000 genes whose absolute correlation with the outcome is largest and then you apply k-means (with k=2) clustering to the dataset with 98 patients and the 1000 selected genes. You find that the clustering produces two groups which have very different survival times using a t-test. Professor Sarkar is delighted and ready to publish. Comment. **Solution**: You are using the data twice! Of course the clusters will have different survival rates!

2. **CLASSIFICATION (8 points)**

In class we considered logistic regression and Gaussian Naive Bayes. In particular, we considered the following types of Gaussian Naive Bayes. We will denote by **GNB-1** the Gaussian Naive Bayes model with two gaussians with means $\mu_1$, $\mu_2$, proportion $\pi$, and covariance matrices $\Sigma_1 \neq \Sigma_2$. We will also denote by **GNB-2** the Gaussian Naive Bayes with means $\mu_1, \mu_2$ and $\Sigma_1 = \Sigma_2$. Because these are Naive Bayes models, $\Sigma_1, \Sigma_2$ are diagonal.

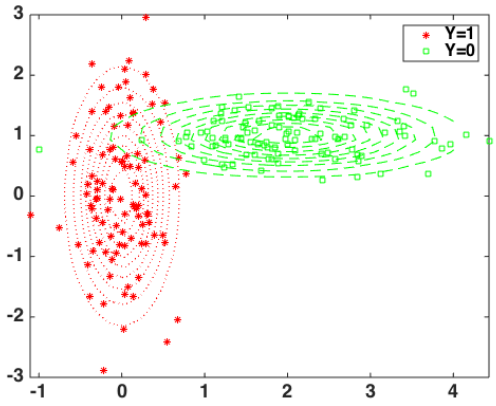
We will denote standard Logistic Regression by **LR**. We will also consider another variant of Logistic Regression **LR-q** where we add features $X_i X_j$ for $1 \leq i \leq j \leq p$. Essentially,

$$\log \frac{P(Y=1|X,\beta)}{P(Y=0|X,\beta)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \sum_{i=1}^{p} \sum_{j=i}^{p} \beta_{ij} X_i X_j$$

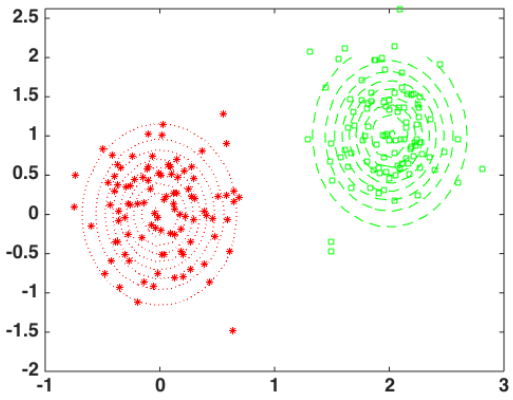
Note that where LR estimates a p+1 dimensional weight vector ($\beta$), LR-q estimates a vector of length $(p+1)(p+2)/2$.

You will need to use the fact that for data generated from two Gaussians with means $\mu_i$ and full covariance matrices $\Sigma_1 \neq \Sigma_2$, the decision boundary is quadratic.

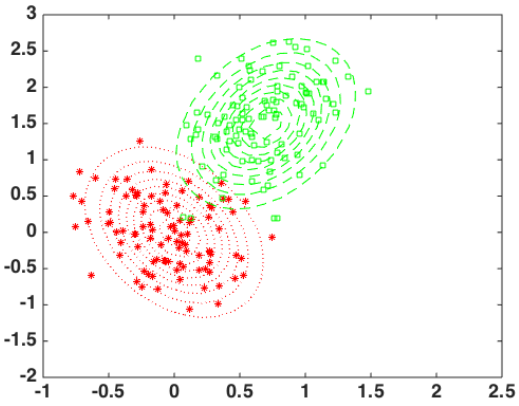
(a) (1+1+2+1 points) Consider the following datasets generated from different sets of pairs of Gaussian distributions. Which models (LR, LR-q, GNB-1, GNB-2) will you use to learn the decision boundaries in each figure? If none of these can do it write "none". (No explanation necessary).



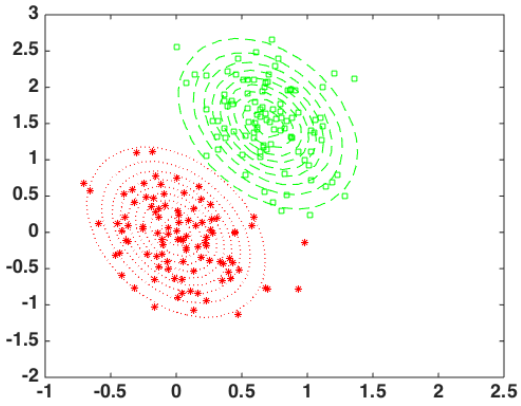

**Solution**: LR-q, GNB-1




**Solution**: all of them

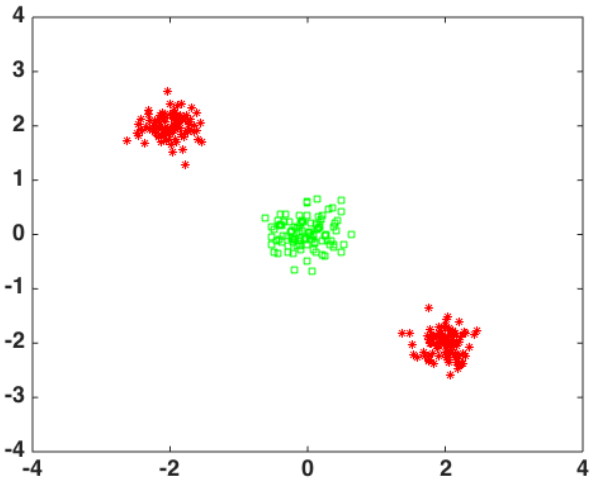



**Solution**: LR-q




**Solution**: LR, LR-q

(b) (3 points) Now consider the following dataset. Write down which models (LR, LR-q, GNB-1, GNB-2) can be used to **correctly classify** the data. If none of these can do it write "none". Are your findings in line with what you learned in class, i.e. LR dominates GNB-2? Explain.




**Solution**: LR-q, GNB-1. Neither LR not GNB-2 can do it since not linear decision boundary. This does not violate what we saw. The problem is difficult for both these methods.