STAT 154 Spring 2019: Final Exam

Instructor: Prof. Bin Yu
May 16, 7 PM–10 PM

Name: ______________________ Student ID: ______________________

Block Number: ________ Row Number: ________ Seat Number: ________

Maximum Points: 100 Time: 2 hours, 50 minutes

Instructions (please read carefully)

- Do not turn the page until you are told do so. **Count and make sure you have 10 sheets of paper.**
- Write your student-id clearly on top of each page.
- **Answer written only in the space provided after the question will be graded.** Do rough work in the extra blank sheets space. There is a help-sheet with the pages provided for rough work.
- **Few multiple choice parts in Q1, Q5 and Q6 may have one or more correct options, and we mention whenever such a question is asked.** *For these parts,* marking all correct options gets full points. Marking any wrong option leads to zero points. Marking a subset of correct options and NO incorrect option gets half of the points.

Pre-Exam Questions

1. What are your favorite restaurants in Berkeley?

China Village, Udupi, La Note

2. Express your feelings for the class.

Overwhelming but very enjoyable.

1 Objective Questions (no justification, 16 parts, 32 pts)

Please darken the corresponding bubble properly. e.g., How often do you like teaching this class?

●Always ○ Sometimes ○ Rarely ○ Never

1.1 Single choice correct

For the next 14 parts, each of them has EXACTLY ONE correct choice.

(i) Which of the following statements is **CORRECT** when the features (columns) in the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ are highly correlated?

● The interpretation for the OLS coefficients becomes difficult.

○ The solution for the lasso becomes unique.

○ The solution for ridge-regression can be non-unique.

○ In such a scenario, using regularization makes sense only if n is much larger than d.

(ii) Select the **INCORRECT** statement related to regularization / bias-variance:

○ ℓ_1-regularization can be used in classification or regression to obtain sparse models.

● Bias-variance trade-off occurs only for regression tasks and does not occur in classification.

○ Using PCA features can provide regularization.

○ The bias of a model increases as the regularization increases.

(iii) Select the **CORRECT** statement.

○ Results from prediction algorithms are always connected to reality in data because they are run on data.

● The dotted red circle in the 3-circle representation of the prediction problem is necessary since the future is almost always different from present data.

○ The 3-partition protocol in the construction of a prediction rule guarantees that the prediction rule will work for future data.

○ Data collection process usually has no-to-little influence on the outcome of a prediction problem.

(iv) Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{1000\times 1000}$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{1000}$. Which of the following commands will take the **LARGEST TIME** to execute in R?

○ `A + B` ○ `A%*%u + B%*%v` ● `(A%*%B) %*% u` ○ `A %*% (B%*%u)`

(v) Suppose that the matrix $\mathbf{X} \in \mathbb{R}^{5\times 3}$ has singular values 3, 2 and 1 and let $\mathbf{u}_i, \mathbf{v}_i, i = 1, 2, 3$ denote its corresponding ordered left and right singular vectors. Then which of the following statements is **CORRECT**?

○ The eigenvalues of the matrix $\mathbf{X}^\top\mathbf{X}$ are 3, 2 and 1.

○ The best 2-rank approximation of the matrix $\mathbf{X}$ (in Frobenius norm) is given by $\mathbf{u}_1\mathbf{v}_1^\top + 2\mathbf{u}_2\mathbf{v}_2^\top$.

○ The first principal component of the matrix $\mathbf{X}$ explains less than 60% of the variance in the data.

● Two of the eigenvalues of the matrix $\mathbf{X}\mathbf{X}^\top$ are 0.

(vi) Which of the following statements is **TRUE** with regards to SVM?

○ As the penalty $C \to \infty$, the soft-margin SVM simplifies to a trivial classifier.

○ As the penalty $C \to 0$, the soft-margin SVM tends towards a hard-margin SVM classifier.

● For linearly separable data, the data points far away from the boundary have no effect on the SVM classifier.

○ The solution of a soft-margin SVM classifier for linearly separable data would be different than that of a hard-margin SVM classifier.

(vii) Which of the following statements is **TRUE** about Adaboost?

○ The main advantage of Adaboost is that training of each tree can be done in parallel.

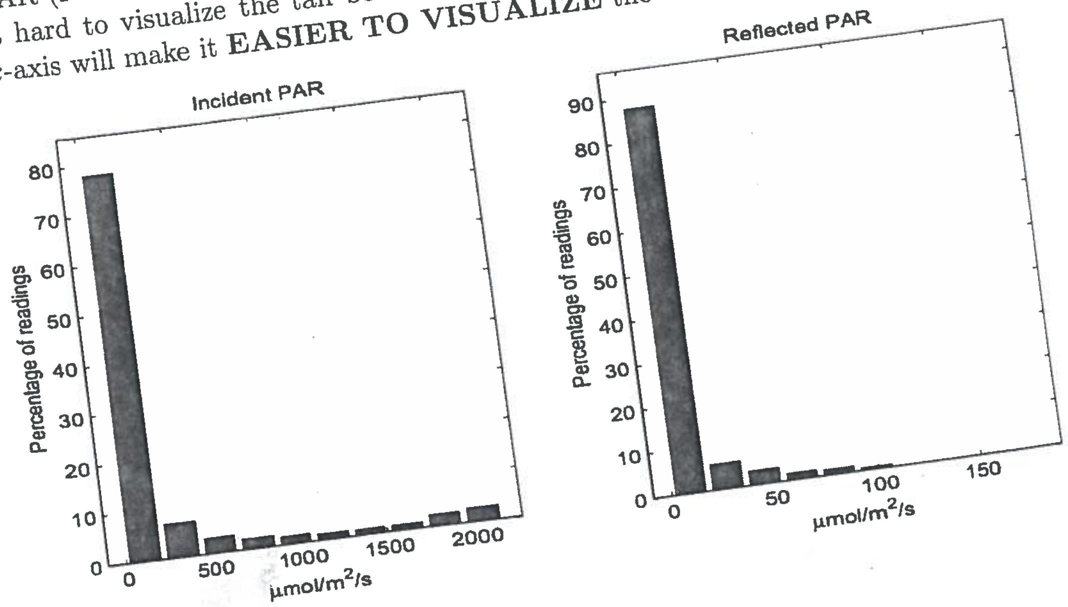
○ In each iteration, new samples are selected from the training data using uniform sampling with replacement.

○ In each iteration, the training samples get re-weighted: the weight of the misclassified points is reduced and that of the correctly classified is increased.

● Adaboost can be seen as a special case of forward stagewise modeling with exponential loss function.

SID:

(viii) Here are two histogram plots from Figure 3(a) of Project 1 on incident and reflected PAR (Photo-synthetically active radiation). These two figures are not good because it is hard to visualize the tail behaviors. Which of the following transformation of the x-axis will make it **EASIER TO VISUALIZE** the tail behavior?



● logarithmic $x \mapsto \log(x)$
○ square $x \mapsto x^2$
○ exponential $x \mapsto \exp x$
○ Cubic $x \mapsto x^3$

(ix) The bias-variance decomposition of a **ridge regression estimator** when compared to ordinary least squares estimator has:

○ higher bias, higher variance
○ smaller bias, smaller variance
○ smaller bias, higher variance
● higher bias, smaller variance

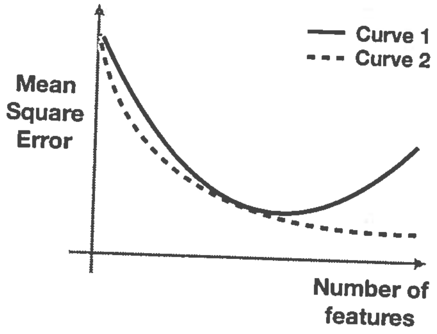
(x) Consider a data-frame "mpg" from R. Which of the following codes will **NOT give the** number of missing values in **each column**?

○ `colSums(is.na(mpg))`
○ `apply(is.na(mpg), 2, sum)`
○ `sapply(mpg, function (x) sum(is.na(x)))`
● `table(is.na(mpg))`

(xi) Consider the Ames Housing data set (in homework 5). The variable column **Overall.Qual** is a categorical variable and contains the customer ratings of the house taking values in $\{1, ..., 10\}$ and we would like to use this variable as a feature in a LASSO model. Then which of the following transformation will ensure that glmnet package for LASSO would treat such a **categorical variable CORRECTLY**?

○ `log(Overall.Qual)`
● `factor(Overall.Qual)`
○ `as.numeric(Overall.Qual)`
○ `scale(Overall.Qual)`

(xii) Consider the following figure that shows the mean-square error (MSE) for a machine learning problem as a function of number of features in the model. Which of the following statements is likely to be **CORRECT**?



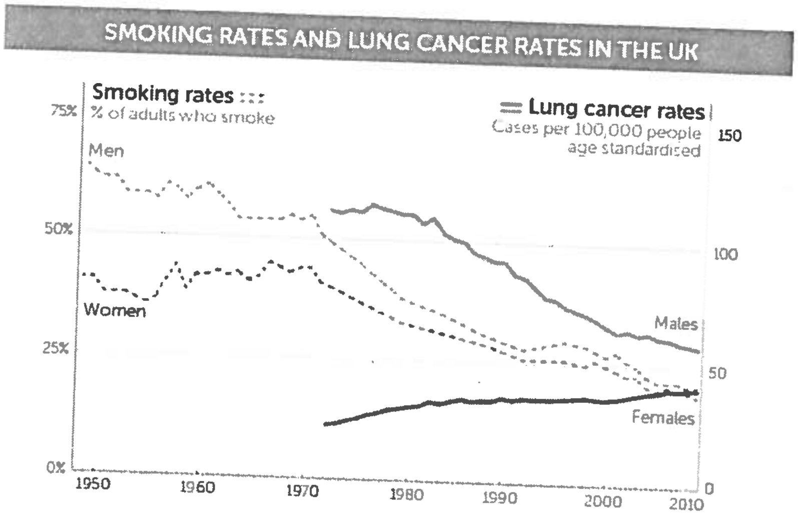
○ Curve 1 denotes the training MSE, and Curve 2 denotes the testing MSE.

○ Curve 1 denotes the training MSE, and Curve 2 denotes the validation MSE.

● Curve 1 denotes the testing MSE, and Curve 2 denotes the training MSE.

○ Such a figure is not possible since both curves are not always decreasing.

(xiii) According to the following figure from UK Cancer Research, which of the following statements is most likely to be **CORRECT**?



○ Smoking causes cancer for men.

○ Before year 2010, smoking was good for women.

● Smoking rate is highly correlated with the lung cancer rate for men.

○ Decrease in smoking rate is likely to increase the lung cancer rate for men.

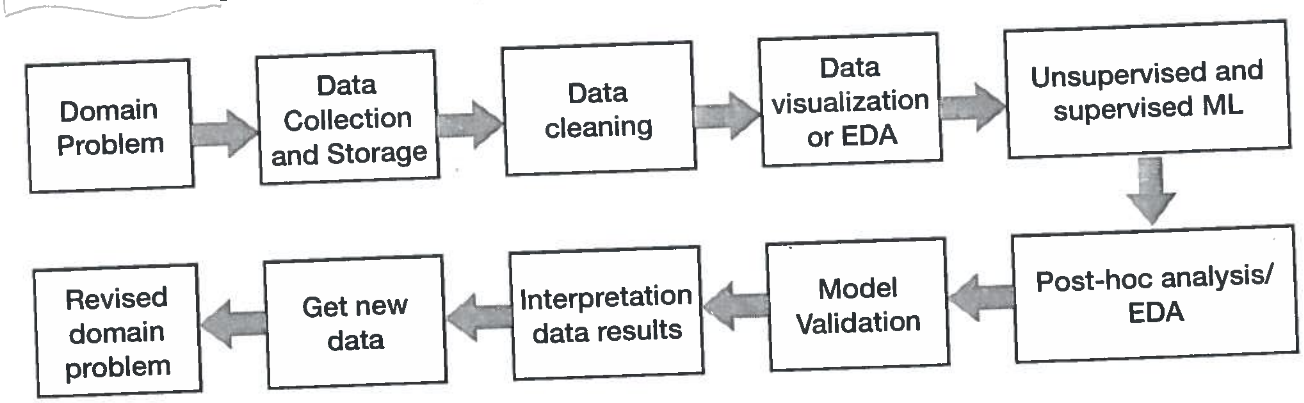
(xiv) Recall the data science life cycle shown in the figure below. If we **SKIP** the "Get new data" step and get back directly to the "Data cleaning" step to try new ways to modeling, which of the following scenario is **likely to HAPPEN**?

○ We obtain a better model to solve the domain problem.

○ Since the data is the same, we enhance the R (representative data collection) notion in PQR-S framing.

● Our model over-fits the existing data after too many post-hoc analyses.

○ Nothing bad will happen.


Domain Problem
Data Collection and Storage
Data cleaning
Data visualization or EDA
Unsupervised and supervised ML
Post-hoc analysis/ EDA
Model Validation
Interpretation data results
Get new data
Revised domain problem

1.2 One or more choice correct

For the next 2 parts, one or more options can be correct.

(xv) Recall the data science life cycle from above (part (xiv)). Which of the steps **REQUIRE** stability analysis?

● Data cleaning

● Data visualization or EDA

● Unsupervised and supervised ML

● Model validation

(xvi) Which of the following way(s) is/are **sensible way(s)** for regularizing a machine learning model, *for both classification and regression*?

● Adding an explicit penalty on the ℓ_1 or ℓ_2 norm of the parameters in the objective function

● Using certain PCA features which explain large fraction of the variance in data

● Early stopping in iterative methods like gradient descent

● Putting additional constraint on the parameters in the optimization problem

2 What is a valid kernel? (10 pts)

Recall that a kernel mapping $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a valid kernel if **either** of the following definitions are met:

(i) There exists a feature map ϕ such that for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, we have $k(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^\top \phi(\mathbf{z})$.

(ii) For all datasets $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the gram matrix K with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ is PSD.

In the following questions, we use the property (i) to show how we can construct new valid kernel mapping or how we can try to verify if a given mapping is a valid kernel mapping.

(a) (2 pts) Let k_1 and k_2 be two valid kernel functions with feature mappings $\phi_1 : \mathbb{R}^d \mapsto \mathbb{R}^p$ and $\phi_2 : \mathbb{R}^d \mapsto \mathbb{R}^p$ such that $k_1(x, y) = \phi_1(x)^\top \phi_1(y)$ and $k_2(x, y) = \phi_2(x)^\top \phi_2(y)$ for all $x, y \in \mathbb{R}^d$. **Show that the mapping $k_+ = k_1 + k_2$ is valid kernel by explicitly constructing its corresponding feature mapping ϕ_+ in terms of ϕ_1 and ϕ_2.**

$$K_+(x,y) = K_1(x,y) + K_2(x,y)$$
$$= \phi_1(x)^T \phi_1(y) + \phi_2(x)^T \phi_2(y)$$
$$= (\phi_1(x) + \phi_2(x))^T (\phi_1(y) + \phi_2(y))$$
$$\Rightarrow \phi_+ = \phi_1 + \phi_2$$

(b) (4 pts) **Show that the mapping $k_* = k_1 \cdot k_2$, that is, $k_*(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) k_2(\mathbf{x}, \mathbf{z})$, is a valid kernel by explicitly constructing its corresponding feature mapping ϕ_* in terms of ϕ_1 and ϕ_2.** *Hint: You might want to start with the case d = 1. If you feel its too tricky, move on and don't remain stuck on it.*

$$K_*(x,y) = K_1(x,y) \; K_2(x,y)$$
$$= \phi_1(x)^T \phi_1(y) \; \phi_2(x)^T \phi_2(y)$$
$$= (\underbrace{\phi_1(x) \circ}_{\uparrow} \phi_2(x))^T (\phi_1(y) \circ \phi_2(y))$$

element wise product.

$$\phi_*(x) = \begin{bmatrix} \phi_{1,1}(x) \cdot \phi_{2,1}(x) \\ \phi_{1,2}(x) \cdot \phi_{2,2}(x) \\ \vdots \\ \phi_{1,p}(x) \cdot \phi_{2,p}(x) \end{bmatrix} \longrightarrow \text{where } \phi_1(x) = \begin{bmatrix} \phi_{1,1}(x) \\ \vdots \\ \phi_{1,p}(x) \end{bmatrix}$$

$$\phi_2(x) = \begin{bmatrix} \phi_{2,1}(x) \\ \vdots \\ \phi_{2,p}(x) \end{bmatrix}$$

(c) (2 pts) Consider the exponential kernel in one dimension $k(x,z) = \exp(xz)$, where $x, z \in \mathbb{R}$. **What is its feature mapping ϕ such that $k(x,z) = \phi(x)^\top \phi(z)$?**

$$e^{xz} = 1 + xz + \frac{x^2 z^2}{2!} + \cdots$$

$$= \underbrace{\left[1, x, \frac{x^2}{\sqrt{2!}}, \frac{x^3}{\sqrt{3!}}, \ldots\right]}_{\phi(x)^T} \underbrace{\begin{bmatrix} 1 \\ z \\ z^2/\sqrt{2!} \\ z^3/\sqrt{3!} \\ \vdots \end{bmatrix}}_{\phi(z)}$$

(d) (3 pts) Now consider the exponential kernel in d-dimensions: $k : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ defined as follows

$$k(\mathbf{x}, \mathbf{z}) = \exp(\mathbf{x}^\top \mathbf{z}), \quad \text{where} \quad \mathbf{x}, \mathbf{z} \in \mathbb{R}^d.$$

Show that this exponential kernel is a valid kernel for any d. Explicit feature mapping is **NOT** required.
Hint: You may directly use the results from any of the previous parts.

Clearly $k(x,z) = e^{x^T z} = e^{x_1 z_1 + x_2 z_2 + \cdots + x_d z_d}$.

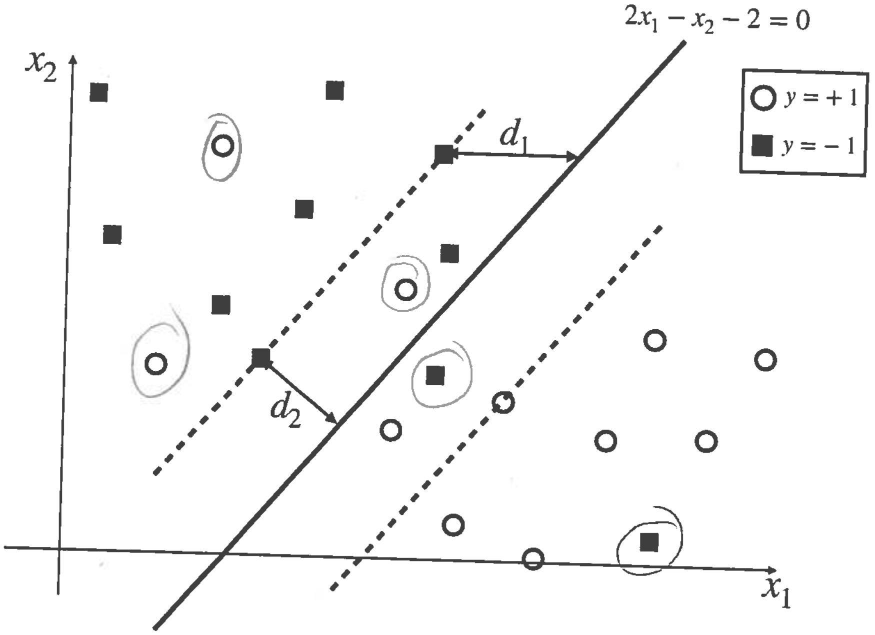
$= K_1(x,z) \cdot K_2(x,z) \ldots K_d(x,z)$

→ where $K_i(x,z) = e^{x_i z_i}$ (where x_i, z_i are i-th coordinate of x, z). is a valid kernel (part(c))

→ Since product of two kernels is valid (part(b)) we can use induction to get that product of 'd' kernels is also valid kernel, and given kernel is expressed above as a product of 'd' valid kernels.

3 SVM and Kernels: Just look and tell. (10 pts)

In the next figure, we plot the results for a soft-margin SVM for binary classification where the classifier is linear and of the form $\mathbf{w}^\top \mathbf{x} + b$. All the training points are plotted. Here the features $\mathbf{x}_i \in \mathbb{R}^2$ and $y_i = +1$ (circle) and $y_i = -1$ (square).



(a) (2 pts) If ξ_i denotes the slack for i-th data point, then **which of the following conditions do all of the support vectors satisfy for a soft-margin SVM?** (If you really want, the help sheet has the soft-margin SVM formulation.)

hard-margin → ○ $y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$

soft-margin → ● $y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1 - \xi_i$

○ $y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 1 - \xi_i$

○ $y_i(\mathbf{w}^\top \mathbf{x}_i + b) < 1 - \xi_i$

(b) (1 pt) In the given figure, **how many points will be incorrectly classified** by the rule $\widehat{y} = \text{sign}(2x_1 - x_2 - 2)$? Just write the number **neatly** since this question will be graded automatically by software.

5 . (points are marked above

(c) (2 pts) Select the **CORRECT** choice for the values d_1, d_2 denoted on the figure. Note that the equation of the solid line is given by $2x_1 - x_2 - 2 = 0$.

○ $d_1 = d_2 = \frac{1}{\sqrt{5}}$

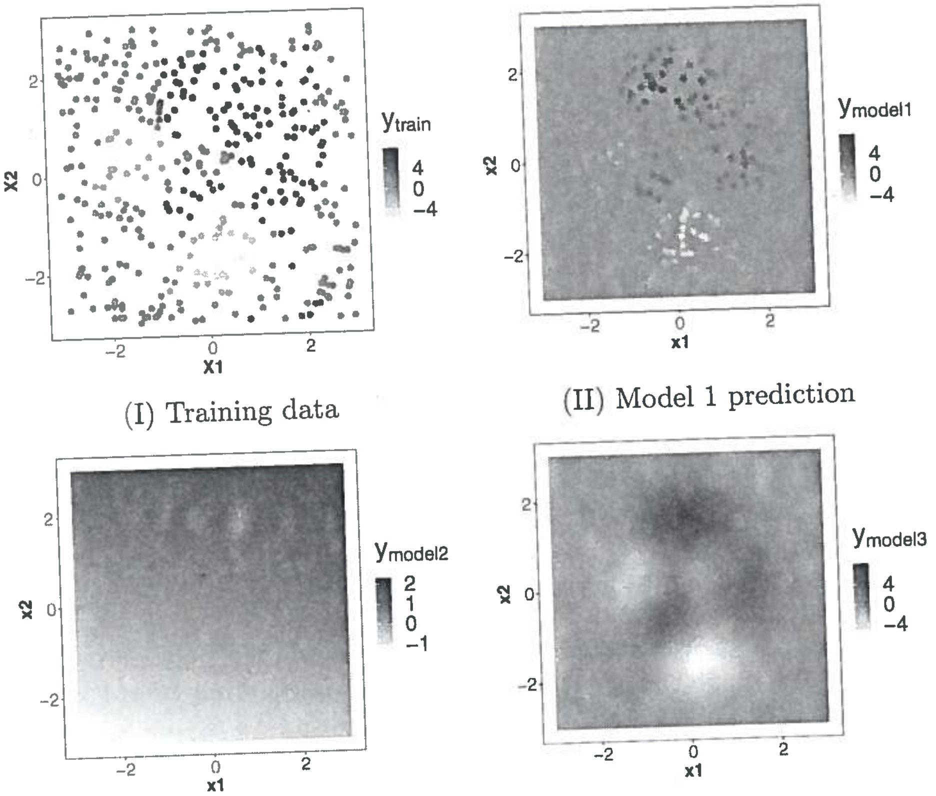
● $d_1 \neq \frac{1}{\sqrt{5}}, d_2 = \frac{1}{\sqrt{5}}$

○ $d_1 = d_2 = \sqrt{5}$

○ $d_1 \neq \sqrt{5}, d_2 = \sqrt{5}$

$\text{margin} = \frac{1}{\|w\|} = \frac{1}{\sqrt{5}}$

For the next 3 parts, we consider a regression task where the samples are coming from a non-linear model $y_i = f(\mathbf{x}_i) + \varepsilon_i$ where ε_i denotes i.i.d. Gaussian noise. In the figure below, we show the noisy training data in panel (I), and the predictions of three different fitted models (one using OLS, two models using Kernel ridge regression) over the entire range of $\mathbf{x}$ (training as well test points) in panels (II), (III) and (IV). *We have not specified which model corresponds to which method.*



(I) Training data

(II) Model 1 prediction

(III) Model 2 prediction

(IV) Model 3 prediction

(d) (2 pts) Among the three models, which one is likely to have the **highest training mean square error (MSE)**, i.e., which model under-fits the training data?

○ Model 1 ● Model 2 ○ Model 3

(e) (2 pts) Which of the model was fitted using **ordinary least squares**?

○ Model 1 ● Model 2 ○ Model 3

(f) (1 pt) If the unknown function was assumed to be smooth, which of the models would you expect to have a **lowest test mean square error** (MSE), i.e., which model would generalize the best?

○ Model 1 ○ Model 2 ● Model 3

4 OLS versus WLS in one dimension (19 pts)

Suppose that we have n data points $\{x_i, y_i\}_{i=1}^n$ generated from the following model:

$$y_i = x_i \beta_* + \varepsilon_i$$

where ε_i are noise variables independent of x_i. We assume that the features $x_i \in \mathbb{R}$ are scalar and fixed. We consider two models to fit the data—ordinary least squares (OLS) and weighted least squares (WLS) with positive weights w_i:

$$\widehat{\beta}_{\text{OLS}} = \arg\min_{\beta} \sum_{i=1}^n (y_i - x_i\beta)^2 \quad \text{and}$$

$$\widehat{\beta}_{\text{WLS}} = \arg\min_{\beta} \sum_{i=1}^n w_i (y_i - x_i\beta)^2.$$

For an estimator $\widehat{\beta}$, we define its mean squared error as $\text{MSE}[\widehat{\beta}] = \mathbb{E}(\widehat{\beta} - \beta_*)^2$ where the expectation is taken over the noise variables.

(a) (1 pt) The estimator $\widehat{\beta}_{\text{OLS}}$ is also the maximum likelihood estimator if the noise variables $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d.

● True ○ False

(b) (1 pt) The estimator $\widehat{\beta}_{\text{OLS}}$ is also the maximum likelihood estimator if the $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ are independent but not necessarily identical (i.e., σ_i's are different).

○ True ● False

(c) (2 pts) Write the expressions for $\widehat{\beta}_{\text{OLS}}$ in terms of $\sum_{i=1}^n x_i y_i$ and $\sum_{i=1}^n x_i^2$; and the expression for $\widehat{\beta}_{\text{WLS}}$ in terms of $\sum_{i=1}^n w_i x_i y_i$, and $\sum_{i=1}^n w_i x_i^2$.

$\widehat{\beta}_{\text{OLS}} =$ $\sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$

$\widehat{\beta}_{\text{WLS}} =$ $\sum_{i=1}^{n} w_i x_i y_i / \sum_{i=1}^{n} w_i x_i^2$

You can directly use these expressions in all the following parts.

(d) (1 pt) The estimator $\widehat{\beta}_{\text{OLS}}$ is an unbiased estimator of β_* if the noise variables $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d.

● True ○ False

Note: For all the following parts, we assume that the noise variables are **independent but not identical**, i.e., $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ where σ_i's are different.

(e) (2 pts) **If $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ are independent but not identical noise variables, is the estimator $\widehat{\beta}_{\text{OLS}}$ an unbiased estimator of β_*? Justify.**

Yes. $\mathbb{E}\,\hat{\beta}_{OLS} = \dfrac{\mathbb{E}\sum x_i y_i}{\sum x_i^2} = \dfrac{\mathbb{E}\sum_{i=1}^{n} x_i(x_i\beta_* + \varepsilon_i)}{\sum x_i^2} = \dfrac{(\sum x_i^2)\,\beta_*}{\sum x_i^2} + \dfrac{\sum x_i\,\cancel{\mathbb{E}\varepsilon_i}}{\sum x_i^2} \xrightarrow{} 0$

$= \beta^*$

(f) (3 pts) **If $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ are independent noise variables and the σ_i's were known, for what values of w_i, the estimator $\widehat{\beta}_{\text{WLS}}$ becomes the MLE estimator of β_*? Is the MLE an unbiased estimator of β_*?**

- likelihood ($\mathcal{L}$) $= \prod_{i=1}^{n} e^{-\frac{(y_i - x_i\beta)^2}{2\sigma_i^2}}$

MLE $= \operatorname{argmax} \mathcal{L} = \operatorname{argmax} \log \mathcal{L}$

$= \operatorname{argmax} \sum_{i=1}^{n} -\dfrac{(y_i - x_i\beta)^2}{2\sigma_i^2}$

$= \operatorname{argmin} \sum_{i=1}^{n} \dfrac{(y_i - x_i\beta)^2}{2\sigma_i^2}$

- Comparing with WLS $w_i = 1/2\sigma_i^2$ or $w_i \propto \dfrac{1}{\sigma_i^2}$

- Yes: $\hat{\beta}_{MLE} = \mathbb{E}\dfrac{\sum_{i=1}^{n} 1/\sigma_i^2\, x_i y_i}{\sum_{i=1}^{n} 1/\sigma_i^2\, x_i^2} = \dfrac{\sum_{i=1}^{n} 1/\sigma_i^2\, x_i^2 \beta_*}{\sum_{i=1}^{n} \sum x_i^2/\sigma_i^2} + 0 = \beta_*$

$(\mathbb{E}\varepsilon_i = 0)$

(g) (3 pts) Compute the mean squared error of the OLS estimator when $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ are independent noise variables, i.e., **derive the expression for MSE($\widehat{\beta}_{\mathbf{OLS}}$)**.
Hint: (i) Recalling the bias-variance trade-off may be useful.
(ii) If $Z_1, \ldots Z_k$ are independent variables, then $Var(a_1 Z_1 + \ldots + a_n Z_n) = \sum_{i=1}^{k} a_i^2\, Var(Z_i)$.

$$MSE(\hat{\beta}) = \overset{\to 0}{\cancel{Bias^2(\hat{\beta})}} + Var(\hat{\beta}) \quad (\hat{\beta}_{OLS} \text{ is unbiased}).$$

$$Var(\hat{\beta}_{OLS}) = Var\left(\frac{\sum_{i=1}^{n} x_i \varepsilon_i}{\sum_{i=1}^{n} x_i^2}\right) = \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \sum_{i=1}^{n} x_i^2 \sigma_i^2$$

$$(\because Var(x_i \varepsilon_i) = x_i^2 \sigma_i^2)$$

(h) (3 pts) Let $w_i = 1/\sigma_i^2$, then **show that**

$$\mathbf{MSE}(\widehat{\beta}_{\mathbf{WLS}}) = \sum_{i=1}^{n} \frac{x_i^2/\sigma_i^2}{\left(\sum_{j=1}^{n} \frac{x_j^2}{\sigma_j^2}\right)^2}.$$

Hints from part (g) might be useful.

$$\text{We have } \mathbb{E}\hat{\beta}_{WLS} = \frac{\sum w_i x_i \mathbb{E} y_i}{\sum w_i x_i^2} = \frac{\sum w_i x_i^2 \beta_*}{\sum w_i x_i^2} = \beta_* \quad (\text{unbiased})$$

$$\therefore MSE(\hat{\beta}) = Var(\hat{\beta}) = Var\left(\frac{\sum w_i x_i \varepsilon_i}{\sum w_i x_i^2}\right) = \frac{1}{(\sum w_i x_i^2)^2} \sum_{i=1}^{n} w_i^2 x_i^2 \sigma_i^2$$

$$w_i = 1/\sigma_i^2 \Rightarrow MSE = \frac{\sum_{i=1}^{n} x_i^2/\sigma_i^2}{\left(\sum_{i=1}^{n} x_i^2/\sigma_i^2\right)^2} \quad ✓$$

$$= \frac{1}{\left(\sum_{i=1}^{n} x_i^2/\sigma_i^2\right)^2}$$

(i) (3 pts) Let $n = 2$ (only two data points). Use the expressions from the previous two parts and **show that when $\sigma_1 \neq \sigma_2$, we have**

$$\mathbf{MSE}(\widehat{\beta}_{\mathbf{WLS}}) < \mathbf{MSE}(\widehat{\beta}_{\mathbf{OLS}})$$

i.e., $\widehat{\beta}_{\mathbf{WLS}}$ is a better estimator than $\widehat{\beta}_{\mathbf{OLS}}$ when the noise variables are not identical. *Although we have given $n = 2$ to make computations easier, you may feel that the part is still tricky. In that case, don't get stuck on it.*

To show

$$\frac{\sum \frac{x_i^2}{\sigma_i^2}}{\left(\sum \frac{x_i^2}{\sigma_i}\right)^2} = \frac{1}{\left(\sum \frac{x_i^2}{\sigma_i^2}\right)} < \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2}$$

$$\iff \left(\sum x_i^2\right)^2 < \left(\sum x_i^2 \sigma_i^2\right)\left(\sum \frac{x_i^2}{\sigma_i^2}\right)$$

$$\|$$

$$\left(\sum (x_i\sigma_i)\left(\frac{x_i}{\sigma_i}\right)^2\right) < \underbrace{\left(\sum x_i^2\sigma_i^2\right)}_{\uparrow \; \|u\|^2} \underbrace{\left(\sum \frac{x_i^2}{\sigma_i^2}\right)}_{\uparrow \; \|v\|^2}$$

If $u = \begin{pmatrix} x_1\sigma_1 \\ \vdots \\ x_n\sigma_n \end{pmatrix}$ $v = \begin{pmatrix} x_1/\sigma_1 \\ \vdots \\ x_n/\sigma_n \end{pmatrix}$

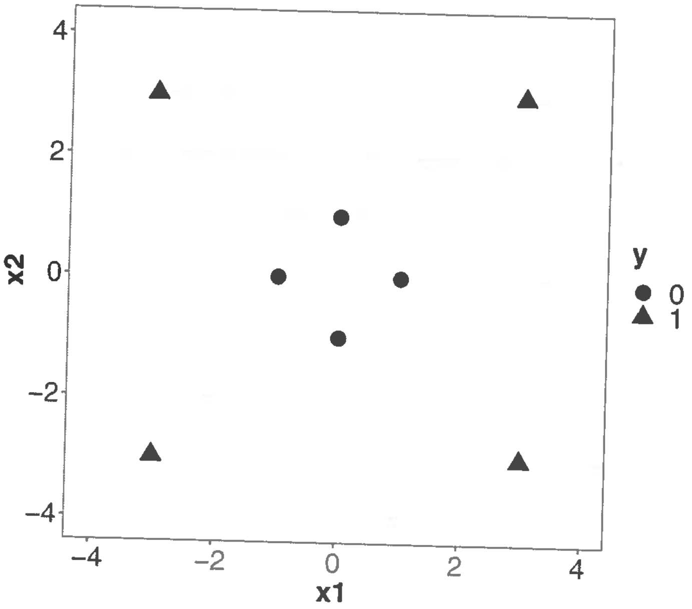
LHS $= (u^T v)^2$

RHS $= \|u\|^2 \; \|v\|^2$

Thus inequality follows from Cauchy Schwarz's inequality

5 Who can classify? (14 pts)

Consider the toy-dataset displayed below. We have 8 data points, with two dimensional feature vector and the labels taking discrete values in $\{0, 1\}$. In short, we have $(\mathbf{x}_i, y_i) \in \mathbb{R}^2 \times \{0, 1\}$ for $i = 1, \ldots, 8$. The scatter plot for the data is shown below.



(a) (1 pt) Which of the following statements is **TRUE** for the dataset shown in the above figure? *Only one correct option.*

○ The data is not-linearly separable and there is no hope to build a good classifier.

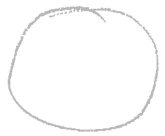
● The data is not-linearly separable and we can use different kernels with different classification methods to build a good classifier.

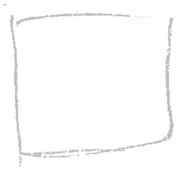
○ Since the data is not-linearly separable, we have to use neural networks to get a good classifier for this dataset.

○ The data is linearly separable.

(b) (2 pts) Select **all the methods** that would achieve 100% training accuracy on this dataset of 8 points with raw features $\mathbf{x}_i$ (no kernel features). *One or more correct options are possible here.*

○ LDA ● QDA ○ Logistic regression ○ SVM ● Decision Tree

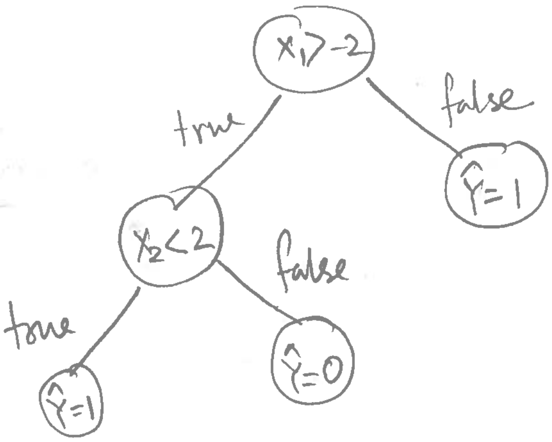


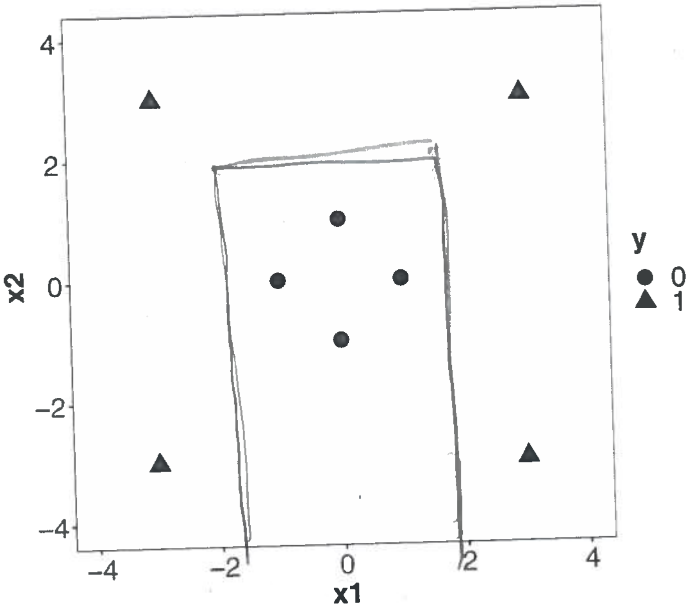


(c) (3 pts) Suppose we use decision trees using ~~decision stumps~~ multiple for classifying this dataset. **Find a decision tree that classifies this dataset as best as possible. Draw its boundary in the figure and write it as a tree. What is its training accuracy?**

(Multiple solutions possible)

100% accuracy

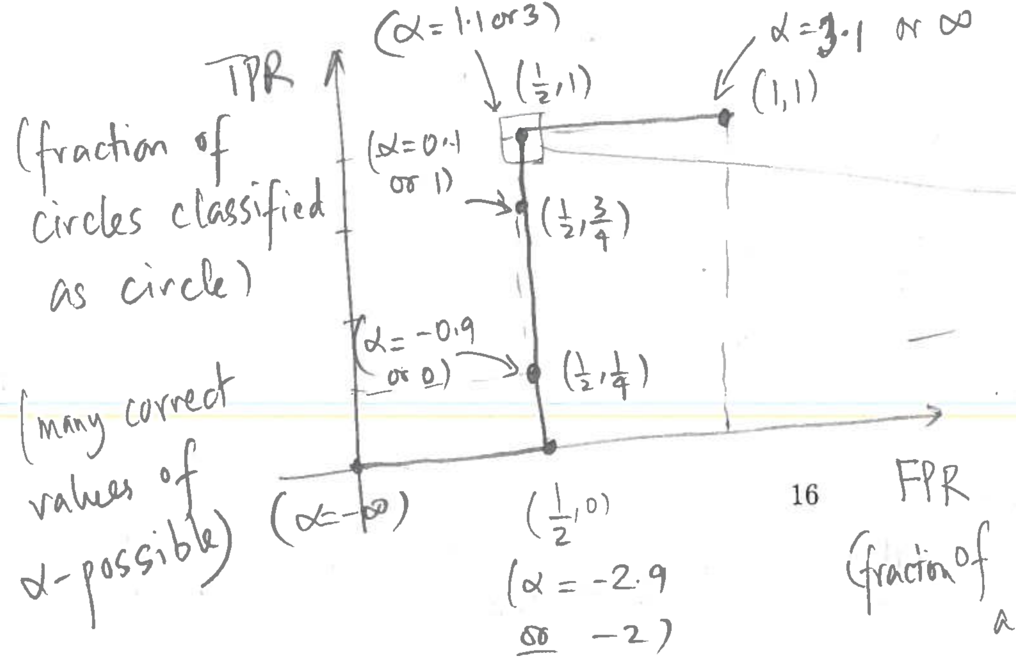




(d) (3 pts) Suppose we want to investigate the ROC curve for prediction of class 0, for the following simple classification rule:

$$\hat{y}_i(\alpha) = \begin{cases} 0 & \text{if} \quad x_{i,1} < \alpha \\ 1 & \text{if} \quad x_{i,1} \geq \alpha \end{cases}$$

Draw the ROC curve for true and false positive rates for this set of functions with respect to the class 0. Can you find a good model from this ROC curve?



→ The best model would be corresponding to

→ TPR = 1, FPR = $\frac{1}{2}$

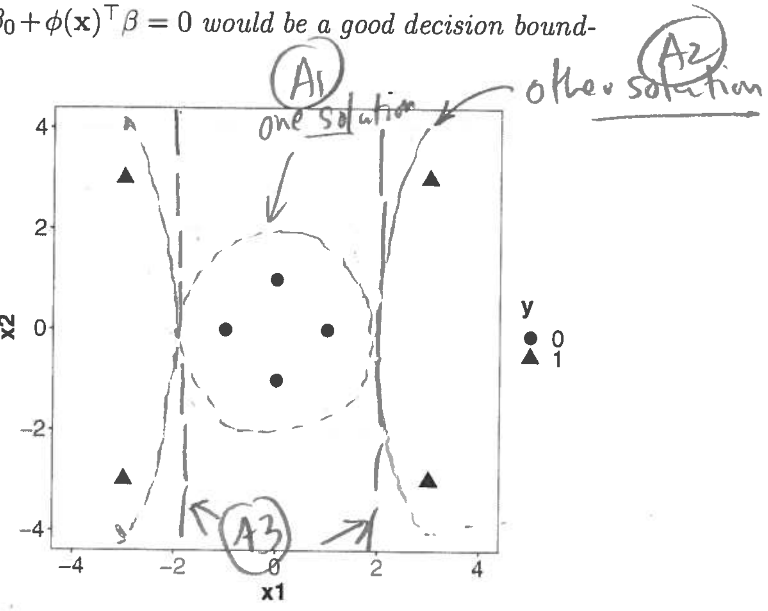
→ May/may not be good enough.

(e) (5 pts) Suppose we use logistic-regression with the kernel $k(\mathbf{x},\mathbf{z}) = (\mathbf{x}^\top\mathbf{z})^2$ but we decide to solve logistic regression using features (i.e., without using the kernel trick) and ℓ_1-regularization to induce sparsity:

$$\min_{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^d} \lambda\|\beta\|_1 + \sum_{i=1}^{8}\left[-y_i(\beta_0+\phi(\mathbf{x}_i)^\top\beta) + \log(1+e^{\beta_0+\phi(\mathbf{x}_i)^\top\beta})\right] \quad (1)$$

where the feature map $\phi : \mathbb{R}^2 \to \mathbb{R}^d$ is chosen corresponding to the quadratic kernel mentioned above. **What is the feature mapping ϕ corresponding to the given kernel $k(\mathbf{x},\mathbf{z}) = (\mathbf{x}^\top\mathbf{z})^2$? Suggest a possible solution $\widehat{\beta}_0, \widehat{\beta}$ to the problem (1) and draw the associated decision boundary approximately in the figure below. Justify briefly.**

Hint: Your solution would be such that $\beta_0+\phi(\mathbf{x})^\top\beta = 0$ would be a good decision boundary on the given figure.



- $$k(x,z) = (x_1z_1 + x_2z_2)^2$$
$$= \underbrace{(x_1^2 \ \ x_2^2 \ \ \sqrt{2}x_1x_2)}_{\phi(x)^T}\begin{pmatrix} z_1^2 \\ z_2^2 \\ \sqrt{2}z_1z_2\end{pmatrix}$$

- $$\phi(x)^T\beta + \beta_0 = 0$$

$$\beta = \begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\end{pmatrix} \Rightarrow x_1^2\beta_1 + x_2^2\beta_2 + \sqrt{2}x_1x_2\beta_3 + \beta_0 = 0$$

ℓ_1-penalty (sparse soln) $\Rightarrow$ one possibility $\beta_3 = 0$

A1 $\Rightarrow x_1^2 + x_2^2 - 2 = 0$ ← circle

another: A2 $x_1^2 - \frac{5}{9}x_2^2 - 4 = 0$ ↖ hyperbola

or other similar solutions

(3, -3)

A3 $\boxed{x_1^2 - 4 = 0}$

6 Do I know the labels or not? (15 pts)

Consider the following data generation mechanism for the samples $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$:

$$Y \sim \text{Bernoulli}(p)$$
$$X|Y = 0 \sim \mathcal{N}(\mu_0, \Sigma_0)$$
$$X|Y = 1 \sim \mathcal{N}(\mu_1, \Sigma_1).$$

MLE has closed form Solutions

(a) (4 pts) Raaz assumes that the covariances are equal $\Sigma_0 = \Sigma_1 = \Sigma$. Let n_0 and $n_1 = n - n_0$ denote the number of samples labeled 0 and 1 respectively for a given set of $n \gg d$ i.i.d. samples $\mathcal{D}_{\text{complete}} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ from the model, **What is the log likelihood of $\mathcal{D}_{\text{complete}}$ (in terms of the parameters $(p, \mu_0, \mu_1, \Sigma)$). Do the MLE estimates for these parameters (or a subset of them) admit closed form expressions with the dataset $\mathcal{D}_{\text{complete}}$?** *If yes, provide the expressions. If the answer is no, then suggest the name of a concrete method/algorithm (with a brief justification, no need to derive the algorithm) that you would use to compute the MLE.*

$\swarrow$(after simplifying)

$$\mathcal{L} = \log p(x_i, y_i, i=1, \dots, n \,|\, \mu_0, \mu_1, p, \Sigma)$$

$$= \sum_{i: Y_i=0} \left[\log p - \frac{(x_i - \mu_0)^T \Sigma^{-1} (x_i - \mu_0)}{2} - C - \frac{1}{2} \log \det \Sigma\right]$$

$$+ \sum_{i: Y_i=1} \left[\log(1-p) - \frac{(x_i - \mu_1)^T \Sigma^{-1} (x_i - \mu_1)}{2} - C - \frac{1}{2} \log \det \Sigma\right].$$

$\Rightarrow$ • $\dfrac{d\mathcal{L}}{dp} = \dfrac{n_0}{p} - \dfrac{n_1}{1-p} = 0 \Rightarrow \boxed{\hat{p} = \dfrac{n_0}{n_0 + n_1} = \dfrac{n_0}{n}}$

• $\dfrac{d\mathcal{L}}{d\mu_0} = 0 \Rightarrow \sum_{Y_i=0} \Sigma^{-1}(x_i - \mu_0) = 0 \Rightarrow \sum_{Y_i=0} (x_i - \mu_0) = 0 \Rightarrow \boxed{\hat{\mu}_0 = \dfrac{\sum_{Y_i=0} x_i}{n_0}}$

(Remember $\nabla_{\mu_0} \mu_0^T A \mu_0 = 2A\mu_0$) (for A symmetric)

$|||^{ly}$ $\boxed{\hat{\mu}_1 = \sum_{Y_i=1} x_i / n_1}$

• $\dfrac{d\mathcal{L}}{d\Sigma^{-1}} = \sum_{Y_i=0} (x_i - \mu_0)(x_i - \mu_0)^T + \sum_{Y_i=1} (x_i - \mu_1)(x_i - \mu_1)^T - n\Sigma = 0$

(TRICKY)

$\Rightarrow \boxed{\hat{\Sigma} = \dfrac{1}{n}\left(\sum_{Y_i=0} (x_i - \mu_0)(x_i - \mu_0)^T + \sum_{Y_i=1} (x_i - \mu_1)(x_i \right.}$

[People not deriving $\hat{\Sigma}$ also get full points
People who derived $\hat{\Sigma}$ get BONUS points].

(b) (3 pts) Assume that p was known and you obtained the MLE estimates $\widehat{\mu}_0, \widehat{\mu}_1, \widehat{\boldsymbol{\Sigma}}$ in the previous part. Then, **what is the decision rule for predicting the y-label for a new test data point with feature vector $\mathbf{x}_{\text{test}}$ if we use the maximizer of the posterior probability of y as our estimate? What is the nature of the decision boundary?**

$$\hat{Y} = \begin{cases} 1 & \text{if } P(Y=1 \mid X = x_{test}, \hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}) > P(Y=0 \mid X = x_{test}, \hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}) \\ 0 & \text{otherwise.} \end{cases}$$

- This classifier is the LDA classifier. Since we assume $\Sigma_0 = \Sigma_1$, The boundary would be linear.

$$P(Y=1 \mid X = x) > P(Y=0 \mid X=x)$$

$$\Leftrightarrow \quad P(y=1)\, P(X=x \mid Y=1) > P(y=0)\, P(X=x \mid Y=0)$$

$$\Leftrightarrow \quad p\, e^{-(x-\mu_q)^T \frac{\Sigma^{-1}}{2} (x-\mu_q)} > (1-p)\, e^{-(x-\mu_0)^T \frac{\Sigma^{-1}}{2} (x-\mu_0)}$$

$$\Leftrightarrow \quad x^T \hat{\Sigma}^{-1} (\hat{\mu}_0 - \hat{\mu}_1) \underset{\hat{Y}=0}{\overset{\hat{Y}=1}{\gtrless}} \log \frac{1-p}{p} + \hat{\mu}_1^T \frac{\hat{\Sigma}^{-1}}{2} \hat{\mu}_1 - \frac{\hat{\mu}_0^T \hat{\Sigma}^{-1} \hat{\mu}_0}{2}$$

after simplifying

(c) (2 pts) Suppose that for the previous two parts, Yuansi objects to the assumption of equal covariance and asks to fit the data with general covariances with no assumption, i.e., $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$ and then use the MAP prediction rule for predicting y labels. Select the **CORRECT** statement (one option) related to the Yuansi's procedure.

● The MLE estimates will admit closed form expressions.

○ The decision boundary will be linear.

○ The decision boundary will be cubic.

(d) (2 pts) Which of the following statements are **certainly TRUE** when comparing the procedures of Raaz and Yuansi from previous parts? (one or more options possible)

○ Yuansi would obtain a better training accuracy.

○ Raaz would obtain a better test accuracy.

○ Raaz would obtain a better training accuracy.

● Yuansi would obtain a larger likelihood on the training data.

(e) (4 pts) Suppose that Raaz accidentally loses all the labels $\{y_i\}_{i=1}^n$ and has access to only the incomplete data $\mathcal{D}_{\text{incomplete}} = \{\mathbf{x}_i\}_{i=1}^n$. **What is the log-likelihood of the dataset $\mathcal{D}_{\text{incomplete}}$ in terms of the parameters $(p, \mu_0, \mu_1, \Sigma)$ with the assumption $\Sigma_0 = \Sigma_1 = \Sigma$? Do the MLE estimates for these parameters (or a subset of them) admit closed form expressions with the dataset $\mathcal{D}_{\text{incomplete}}$?** *If yes, provide the expressions. If the answer is no, then suggest the name of a concrete method/algorithm (with a brief justification, no need to derive the algorithm) that you would use to compute the MLE.*

$$p(x_i) = p \frac{e^{-(x_i-\mu_0)^T \frac{\Sigma^{-1}}{2} (x_i-\mu_0)}}{(\sqrt{2\pi})^d \det(\Sigma)} + (1-p) \frac{e^{-(x_i-\mu_1)^T \frac{\Sigma^{-1}}{2} (x_i-\mu_1)}}{(\sqrt{2\pi})^d \sqrt{\det \Sigma}}$$

$$\Rightarrow \log p(x_1, \dots, x_n | \mu_0, \mu_1, \Sigma)$$

$$= \sum_{i=1}^n \left[\log \left\{ p\, e^{-(x_i-\mu_0)^T \frac{\Sigma^{-1}}{2} (x_i-\mu_0)} + (1-p)\, e^{-(x_i-\mu_1)^T \frac{\Sigma^{-1}}{2} (x_i-\mu_1)} \right\} - \log\left[(\sqrt{2\pi})^d \sqrt{\det \Sigma}\right] \right]$$

- No, This is a mixture model and does not have closed form MLE. We can use EM to estimate MLE.