STA314: Statistical Methods for Machine Learning I

Midterm Exam 2 – LEC0101

Problem 1 (3 pts)

In this problem we will derive the regression function of a linear spline with one knot. Let $X \in \mathbb{R}$ be a one-dimensional feature variable and $Y \in \mathbb{R}$ be the response. To fit a linear spline with one knot at $X = x_0$, we start with the following piecewise linear regression function

$$f(X) = \begin{cases} \alpha_0 + \alpha_1 X, & \text{if } X < x_0 \\ \alpha_2 + \alpha_3 X, & \text{if } X \geq x_0 \end{cases} \tag{1}$$

1. (1 pt) State the requirement of linear splines at the knot $X = x_0$. Derive its induced constraint on the coefficients $\alpha_0, \alpha_1, \alpha_2$ and α_3.

   SOLUTION: Requires continuity and induces the constraint $\alpha_0 + \alpha_1 x_0 = \alpha_2 + \alpha_3 x_0$, or

   $$\alpha_2 - \alpha_0 = (\alpha_1 - \alpha_3) x_0.$$

2. (2 pts) Prove that under the constraint you derived in part 1,

$$f(X) = \alpha_0 + \alpha_1 X + (\alpha_3 - \alpha_1)(X - x_0)_+.$$

For any $a \in \mathbb{R}$, we write $(a)_+ = \max\{a, 0\}$.

SOLUTION:

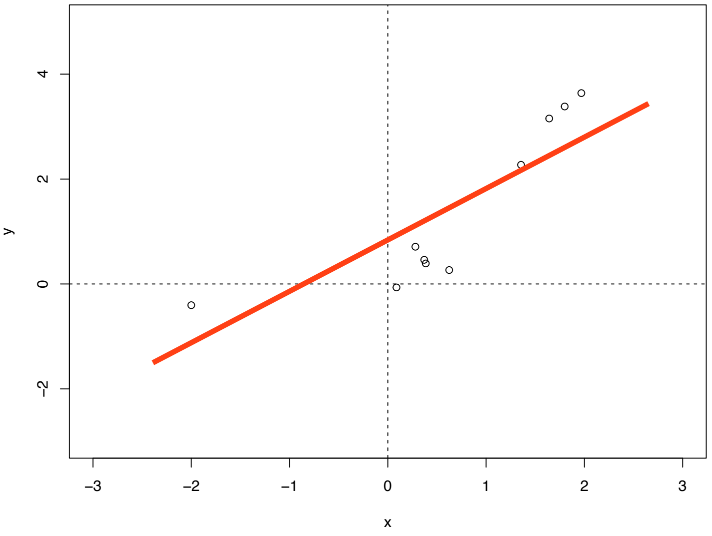
$$\begin{aligned} f(X) &= \alpha_0 1\{X < x_0\} + \alpha_1 X 1\{X < x_0\} + \alpha_2 1\{X \geq x_0\} + \alpha_3 X 1\{X \geq x_0\} \\ &= \alpha_0 + \alpha_1 X + (\alpha_2 - \alpha_0) 1\{X \geq x_0\} + (\alpha_3 - \alpha_1) X 1\{X \geq x_0\} \\ &= \alpha_0 + \alpha_1 X + (\alpha_3 - \alpha_1)(X - x_0) 1\{X \geq x_0\} \end{aligned}$$

The last step uses part 1.

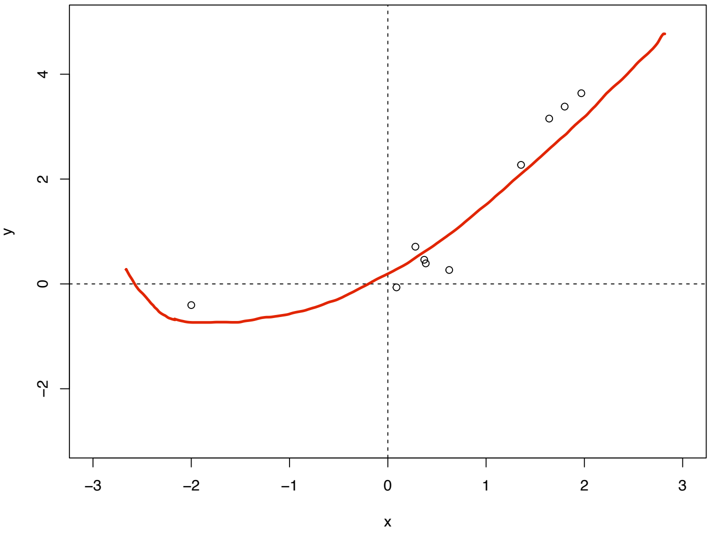
Problem 2 (13 points)

Suppose we observe 10 data points (x_i, y_i) for $1 \leq i \leq 10$ as shown below.

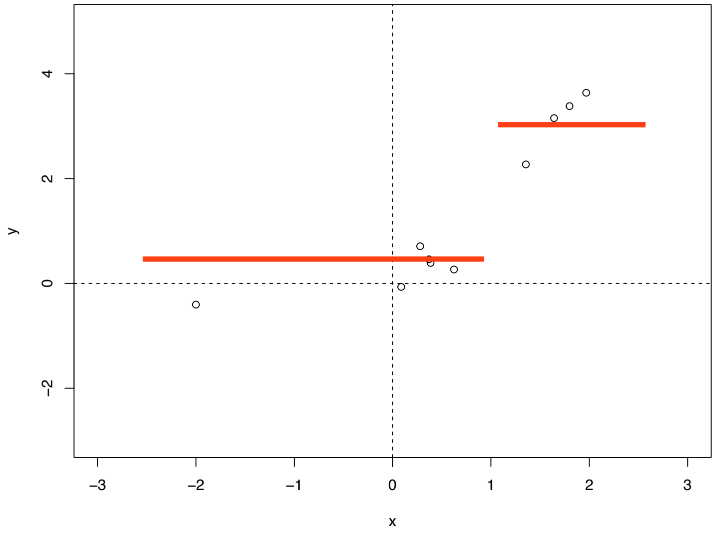
1. (1 pt) If we use OLS to fit a linear regression between y_i and x_i, sketch the fitted prediction line for all $-2.5 < x < 2.5$.



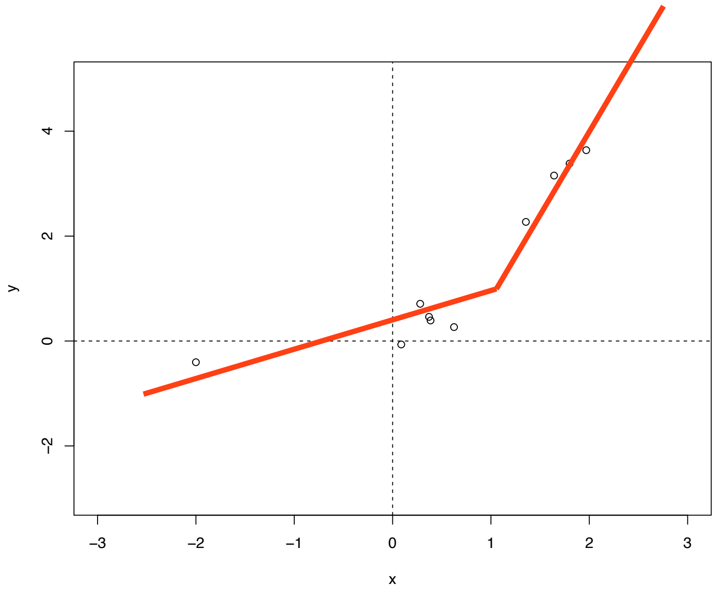
2. (1 pt) If we use OLS to fit a polynomial regression of order 2 between y_i and x_i, sketch the fitted prediction line for all $-2.5 \leq x \leq 2.5$.



3. (1 pt) If we use OLS to fit a step-wise regression between y_i and x_i at the knot $x = 1$, sketch the fitted prediction line for all $-2.5 \leq x \leq 2.5$.



4. (1 pt) If we use OLS to fit a linear spline between y_i and x_i at the knot $x = 1$, sketch the fitted prediction line for all $-2.5 \leq x \leq 2.5$.

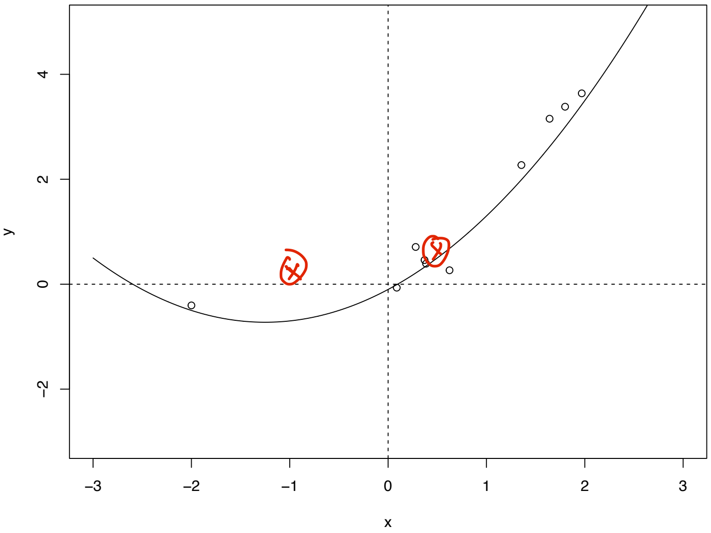


5. (3 pts) For each pair of the four fitted models above, compare their training MSEs (e.g., for model 1 vs model 2, your answer should be one of $\text{MSE}_1 \geq \text{MSE}_2$, $\text{MSE}_1 \leq \text{MSE}_2$ or “there is not enough information to tell”.)

   SOLUTION:

   - model 1 vs model 2: $\text{MSE}_1 \geq \text{MSE}_2$
   - model 1 vs model 3: not able to tell
   - model 1 vs model 4: $\text{MSE}_1 \geq \text{MSE}_4$
   - model 2 vs model 3: not able to tell
   - model 2 vs model 4: not able to tell
   - model 3 vs model 4: $\text{MSE}_3 \geq \text{MSE}_4$

6. (3 pts) Indicate in the figure the position of predicted values at $x = -1$ and $x = 0.5$ by using K-nearest neighbor with $K = 3$. Suppose the solid line is the true regression function. Comment on how you expect the prediction accuracy at a given point to depend on its neighbors.



SOLUTION: The closer the neighbors are, the more accurate.

7. (2 pts) Suppose we have measurement of two features X_1 and X_2. We would like to use a generalized additive model (GAM) to predict the response Y where we use a step-wise basis of X_1 with knot $x_1 = 1$ and a linear spline basis of X_2 with knot $x_2 = 0$. State your model.

   SOLUTION:

$$Y = \beta_0 + \beta_1 1\{X_1 \leq 1\} + \beta_2 X_2 + \beta_3 (X_2 - 1)_+ + \varepsilon.$$

   or simply

$$Y \sim 1 + 1\{X_1 \leq 1\} + X_2 + (X_2 - 1)_+.$$

   There are other valid basis choices of X_1 and X_2.

8. (1 pt) State at least two procedures which can be used to fit the above GAM model.

   SOLUTION: Possible options are OLS, ridge and lasso.

Problem 3 (14 points)

Consider the following 6 data points of two features $X = (X_1, X_2)$ and a three-class label $Y \in \{0, 1, 2\}$. Each row corresponds to one data point.

X_1	X_2	Class Y
-1	1	0
-0.5	2	0
0	3	~~1~~ 2
0.5	2	~~1~~ 2
1	0	~~2~~ 1
1.5	0.5	~~2~~ 1

Suppose we aim to train a multi-class logistic regression classifier with Class 0 chosen as the baseline, that is, for any $\mathbf{x} = (x_1, x_2)$, we assume

$$\log \frac{\mathbb{P}(Y = 1 \mid X = \mathbf{x})}{\mathbb{P}(Y = 0 \mid X = \mathbf{x})} = \beta_0^{(1)} + \beta_1^{(1)} x_1 + \beta_2^{(1)} x_2,$$

$$\log \frac{\mathbb{P}(Y = 2 \mid X = \mathbf{x})}{\mathbb{P}(Y = 0 \mid X = \mathbf{x})} = \beta_0^{(2)} + \beta_1^{(2)} x_1 + \beta_2^{(2)} x_2.$$

Here $\boldsymbol{\beta}^{(1)} = (\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}) \in \mathbb{R}^3$ and $\boldsymbol{\beta}^{(2)} = (\beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}) \in \mathbb{R}^3$ are the unknown coefficients.

1. (3 pts) Derive the expression of $\mathbb{P}(Y = k \mid X = \mathbf{x})$ for all $k \in \{0, 1, 2\}$ in terms of $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$ and $\mathbf{x}$.

   SOLUTION:

$$\mathbb{P}(Y = 0 \mid X = \mathbf{x}) = \frac{1}{1 + e^{\beta_0^{(1)}+\beta_1^{(1)}x_1+\beta_2^{(1)}x_2} + e^{\beta_0^{(2)}+\beta_1^{(2)}x_1+\beta_2^{(2)}x_2}}$$

$$\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \frac{e^{\beta_0^{(1)}+\beta_1^{(1)}x_1+\beta_2^{(1)}x_2}}{1 + e^{\beta_0^{(1)}+\beta_1^{(1)}x_1+\beta_2^{(1)}x_2} + e^{\beta_0^{(2)}+\beta_1^{(2)}x_1+\beta_2^{(2)}x_2}}$$

$$\mathbb{P}(Y = 2 \mid X = \mathbf{x}) = \frac{e^{\beta_0^{(2)}+\beta_1^{(2)}x_1+\beta_2^{(2)}x_2}}{1 + e^{\beta_0^{(1)}+\beta_1^{(1)}x_1+\beta_2^{(1)}x_2} + e^{\beta_0^{(2)}+\beta_1^{(2)}x_1+\beta_2^{(2)}x_2}}$$

2. (2 pts) The MLE can be used to estimate the coefficients $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$, and is defined as

$$\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)} = \underset{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}}{\arg\max}\ \ell(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}).$$

Here $\ell(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})$ is the log-likelihood function. In practice, people also consider the following regularized estimator

$$\underset{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}}{\arg\max} \left\{ \ell(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}) - \lambda \left[\left(\beta_1^{(1)}\right)^2 + \left(\beta_2^{(1)}\right)^2 + \left(\beta_1^{(2)}\right)^2 + \left(\beta_2^{(2)}\right)^2 \right] \right\}$$

where $\lambda > 0$ is some regularization parameter.

Explain the effect of regularization, specifically, how would you expect λ to affect the resulting estimator, as well as the model complexity?

SOLUTION: For larger λ, the magnitude of the estimated coefficients get smaller, the model complexity gets reduced.

3. (3 pts) Suppose we use the following estimates of $\boldsymbol{\beta}^{(1)} = (\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)})$ and $\boldsymbol{\beta}^{(2)} = (\beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)})$:

$$\widehat{\boldsymbol{\beta}}^{(1)} = (1, 1, -1), \qquad \widehat{\boldsymbol{\beta}}^{(2)} = (-1, 1, 1). \tag{2}$$

State the interpretation of both $\widehat{\beta}_1^{(1)}$ and $\widehat{\beta}_1^{(2)}$. Compute the predicted probabilities for each label class for the observation $\mathbf{x} = (1, 0)$.

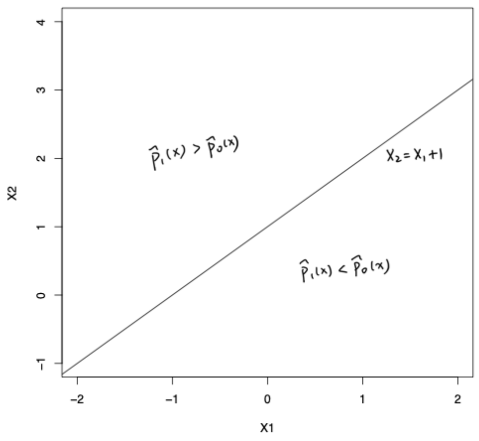
(Your answer may contain terms such as e, e^{-1}, e^2, $e^{-2}, \ldots$)

SOLUTION: For any unit increment of x_1 with the value of x_2 fixed, $\widehat{\beta}_1^{(1)}$ represents the change in the log odds of $\mathbb{P}(Y = 1 \mid X = \mathbf{x})$ relative to $\mathbb{P}(Y = 0 \mid X = \mathbf{x})$ while $\widehat{\beta}_1^{(2)}$ represents the change in the log odds of $\mathbb{P}(Y = 2 \mid X = \mathbf{x})$ relative to $\mathbb{P}(Y = 0 \mid X = \mathbf{x})$.

The predicted probabilities are

$$\mathbb{P}(Y = 0 \mid X = \mathbf{x}) = \frac{1}{2 + e^2}$$
$$\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \frac{e^2}{2 + e^2}$$
$$\mathbb{P}(Y = 2 \mid X = \mathbf{x}) = \frac{1}{2 + e^2}.$$

4. (1 pt) By using the estimates in equation (2), draw the line in a 2-dimensional space (with x-axis indicating the value of X_1 and y-axis indicating the value of X_2) such that $\widehat{p}_1(\mathbf{x}) := \mathbb{P}(Y = 1 \mid X = \mathbf{x}) < \widehat{p}_0(\mathbf{x}) := \widehat{\mathbb{P}}(Y = 0 \mid X = \mathbf{x})$ on one side and $\widehat{p}_1(\mathbf{x}) > \widehat{p}_0(\mathbf{x})$ on the other side. E.g.



Give the mathematical expression of this line and indicate which class has larger probability on each side.

$\hat{p}_1 < \hat{p}_0$

$\Leftrightarrow \frac{\hat{p}_1}{\hat{p}_0} < 1$

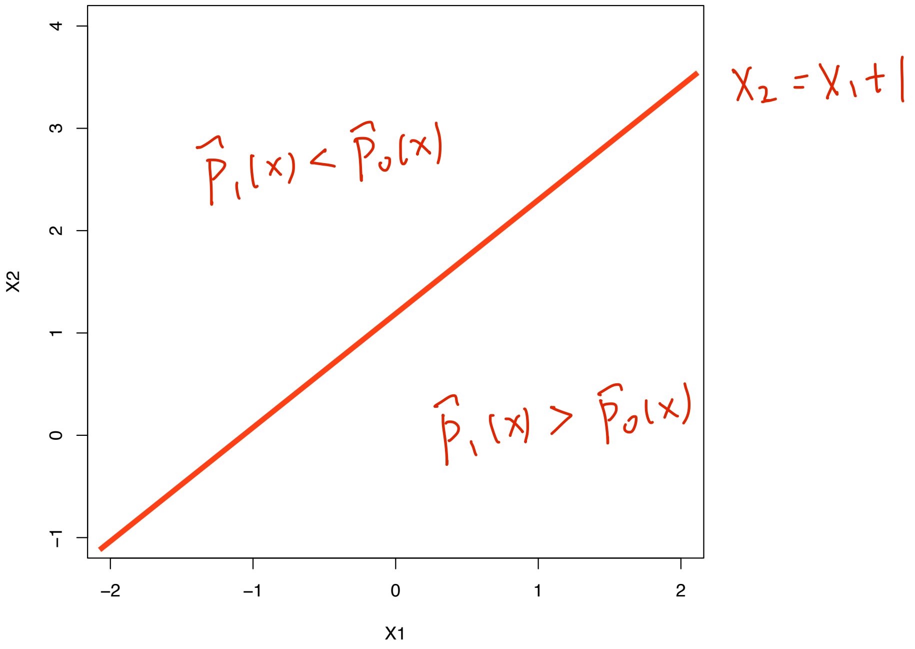
$\Leftrightarrow \log \frac{\hat{p}_1}{\hat{p}_0} < 0$

$\Leftrightarrow \hat{\beta}_0^{(1)} + \hat{\beta}_1^{(1)} x_1 + \hat{\beta}_2^{(1)} x_2 < 0$

Plugging in,

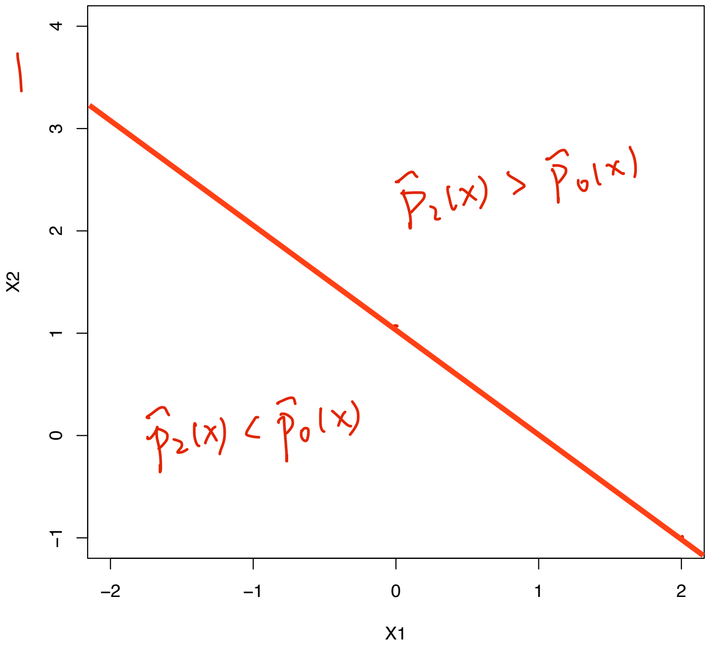
$1 + x_1 - x_2 < 0$

$\Leftrightarrow x_2 > x_1 + 1$



5. (1 pt) Repeat the previous part between $\widehat{p}_2(\mathbf{x}) := \widehat{\mathbb{P}}(Y = 2 \mid X = \mathbf{x})$ and $\widehat{p}_0(\mathbf{x})$. (Draw the line, state its mathematical expression and indicate which class has larger probability on each side)

$X_2 = -X_1 + 1$



$$\hat{p}_2 < \hat{p}_0 \iff \frac{\hat{p}_2}{\hat{p}_0} < 1 \iff \log \frac{\hat{p}_2}{\hat{p}_0} < 0 \iff \hat{\beta}_0^{(2)} + \hat{\beta}_1^{(2)} x_1 + \hat{\beta}_2^{(2)} x_2 < 0$$

Plugging in,

$$-1 + x_1 + x_2 < 0 \iff x_2 < 1 - x_1$$

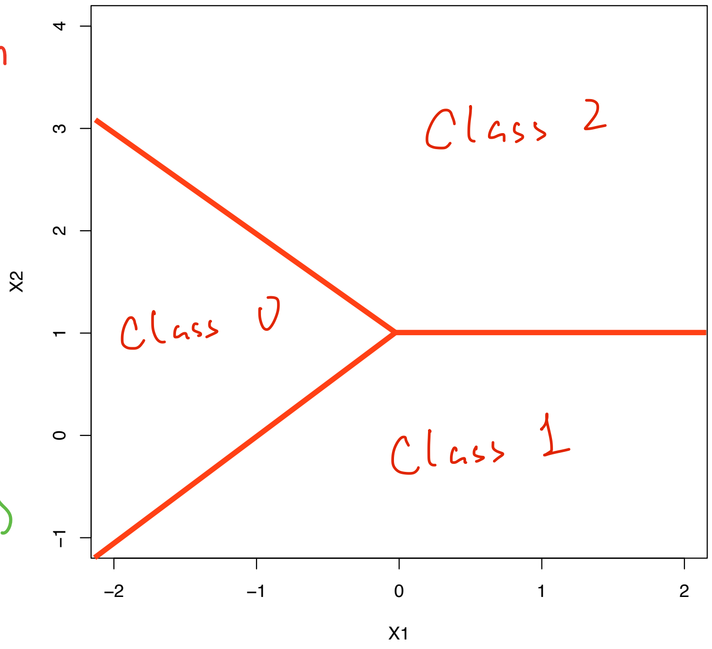
6. (1 pt) Based on the previous two questions, indicate the areas in which the predicted label is Class 0. Similarly, indicate the other two areas where the predicted labels are Class 1 and 2, respectively.

We know boundaries for $\hat{p}_1$ vs. $\hat{p}_0$ and $\hat{p}_2$ vs. $\hat{p}_0$ from previous parts. Just missing $\hat{p}_1$ vs. $\hat{p}_2$.

By Part (1),

$$\frac{\hat{p}_1}{\hat{p}_2} = e^{x^T\hat{\beta}^{(1)} - x^T\hat{\beta}^{(2)}} < 1$$

Here we wrote $x=(1,x_1,x_2)$ for convenience.



$X_2 = 1$

$$\Leftrightarrow \log\frac{\hat{p}_1}{\hat{p}_2} = x^T\hat{\beta}^{(1)} - x^T\hat{\beta}^{(2)} < 0$$

$$\Leftrightarrow x^T\hat{\beta}^{(1)} < x^T\hat{\beta}^{(2)}$$

Plugging in,

$$1+x_1-x_2 < -1+x_1+x_2$$

$$\Leftrightarrow 2x_2 > 2$$

$$\Leftrightarrow x_2 > 1$$

X_1	X_2	Class Y	Predicted
-1	1	0	0
-0.5	2	0	2
0	3	~~1~~ 2	2
0.5	2	~~1~~ 2	2
1	0	~~2~~ 1	1
1.5	0.5	~~2~~ 1	1

Error rate is $\frac{1}{6}$

7. (3 pts) State the predicted labels for each of the training data points (you can either base on the figure in previous part or do the computation directly). Calculate the training error rate.

See table and error rate calculation above.

Problem 4 (10 points)

Multinomial logit model (MLM) is another popular model for multi-class classification problem. Imagine a study where individuals are asked to choose their preferred product among a list of $(K+1)$ items. For each product, we have measurement of its attributes. Here we only consider one attribute, such as price. The prices of each product are $x_0, x_1, \ldots, x_K$. The MLM assumes that the customer makes their choice Y according to

$$\log \frac{\mathbb{P}(Y=k)}{\mathbb{P}(Y=0)} = \beta_0^* + \beta_1^* x_k, \quad k \in \{1, \ldots, K\}.$$

Product 0 is chosen as the baseline. We write $Y = k$ if the customer chooses product k. The unknown coefficients β_0^* and β_1^* represent the customer's "taste" on price. Suppose we observe n i.i.d. choices $y_1, \ldots, y_n$ of a chosen customer according to the above model.

1. (2 pts) Show that

$$\mathbb{P}(Y=0) = \frac{1}{1+\sum_{k=1}^{K} e^{\beta_0^*+\beta_1^* x_k}}$$

$$\mathbb{P}(Y=k) = \frac{e^{\beta_0^*+\beta_1^* x_k}}{1+\sum_{k=1}^{K} e^{\beta_0^*+\beta_1^* x_k}}, \qquad \text{for all } k \in \{1,\ldots,K\}.$$

SOLUTION: By definition, for all $1 \leq k \leq K$,

$$\mathbb{P}(Y=k) = e^{\beta_0^*+\beta_1^* x_k}\mathbb{P}(Y=0).$$

Since

$$1 = \sum_{k=1}^{K} \mathbb{P}(Y=k) + \mathbb{P}(Y=0) = \mathbb{P}(Y=0)\left(e^{\beta_0^*+\beta_1^* x_k}+1\right),$$

we have

$$\mathbb{P}(Y=0) = \frac{1}{1+e^{\beta_0^*+\beta_1^* x_k}}.$$

The other claim follows immediately.

2. (3 pts) Let

$$n_k = \sum_{i=1}^{n} 1\{y_i = k\}, \quad \text{for all } k \in \{0, 1, \ldots, K\}.$$

Prove that the log-likelihood function at any (β_0, β_1) is

$$\ell(\beta_0, \beta_1) = \sum_{k=1}^{K} n_k(\beta_0 + \beta_1 x_k) - n \log\left(1 + \sum_{k=1}^{K} e^{\beta_0 + \beta_1 x_k}\right). \tag{3}$$

SOLUTION: The likelihood of y_1 is

$$L(\beta_0, \beta_1; y_1) = \prod_{k=0}^{K} \mathbb{P}(y_1 = k)^{1\{y_i = k\}}$$

so that the log-likelihood of $y_1, \ldots, y_n$ is

$$\begin{aligned}
&\ell(\beta_0, \beta_1) \\
&= \sum_{i=1}^{n} \sum_{k=0}^{K} 1\{y_i = k\} \log\left[\mathbb{P}(y_1 = k)\right] \\
&= \sum_{i=1}^{n} 1\{y_i = 0\} \left[-\log\left(1 + \sum_{k=1}^{K} \exp(\beta_0 + \beta_1 x_k)\right)\right] \\
&\quad + \sum_{i=1}^{n} \sum_{k=1}^{K} 1\{y_i = k\} \left[\beta_0 + \beta_1 x_k - \log\left(1 + \sum_{k=1}^{K} \exp(\beta_0 + \beta_1 x_k)\right)\right] \\
&= \sum_{k=1}^{K} n_k(\beta_0 + \beta_1 x_k) - n \log\left(1 + \sum_{k=1}^{K} \exp(\beta_0 + \beta_1 x_k)\right).
\end{aligned}$$

(You may use this blank page to continue your answer.)

3. (2 pts) Suppose we know $\beta_0^* = 0$ and we only maximize the log-likelihood function $\ell(\beta_1) := \ell(\beta_0 = 0, \beta_1)$ in (3) over $\beta_1 \in \mathbb{R}$ to compute the MLE of β_1^*. Starting from a given initialization $\widehat{\beta}_1^{(0)}$ with a given step size α, state the gradient descent iterates for computing the MLE of β_1^*. (You need to derive the expression of gradient)

SOLUTION: Write

$$p_0(\beta_0, \beta_1) = \frac{1}{1 + \sum_{k=1}^K e^{\beta_0 + \beta_1 x_k}}$$
$$p_k(\beta_0, \beta_1) = \frac{e^{\beta_0 + \beta_1 x_k}}{1 + \sum_{k=1}^K e^{\beta_0 + \beta_1 x_k}}, \quad k \in \{1, \dots, K\}.$$

Since

$$\begin{aligned}\frac{\partial \ell(\beta_0, \beta_1)}{\partial \beta_1} &= \sum_{k=1}^K n_k x_k - n \frac{\sum_{k=1}^K e^{\beta_0 + \beta_1 x_k} \ x_k}{1 + \sum_{k=1}^K e^{\beta_0 + \beta_1 x_k}} \\ &= \sum_{k=1}^K \left[n_k - n \ p_k(\beta_0, \beta_1)\right] x_k,\end{aligned}$$

the GD of $\widehat{\beta}_1^{(t)}$ follows as

$$\widehat{\beta}_1^{(t+1)} = \widehat{\beta}_1 - \alpha \sum_{k=1}^K \left[n_k - n \ p_k(0, \widehat{\beta}_1^{(t)})\right] x_k.$$

Specifically,

$$p_k(0, \widehat{\beta}_1^{(t)}) = \frac{e^{\beta_1 x_k}}{1 + \sum_{k=1}^K e^{\beta_1 x_k}}.$$

(You may use this blank page to continue your answer.)

4. (3 pts) Suppose $K = 1$. Prove that the negative log-likelihood, $-\ell(\beta_1)$, in the previous subquestion is a convex function of β_1. Reason whether or not the MLE of β_1 can be computed via the gradient descent you derived above with a suitable step size.
(Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is said to be convex if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(x)$ for all $x, y \in \mathbb{R}$ and all $\lambda \in [0,1]$. A sufficient condition of $f(x)$ being convex is $f''(x) \geq 0$ for all x.)

SOLUTION: From previous part,

$$-\frac{\partial^2 \ell(\beta_0, \beta_1)}{\partial \beta_1^2} = n \left\{ \frac{\sum_{k=1}^K x_k^2 e^{\beta_0 + \beta_1 x_k}}{1 + \sum_{k=1}^K e^{\beta_0 + \beta_1 x_k}} - \left(\frac{\sum_{k=1}^K x_k e^{\beta_0 + \beta_1 x_k}}{1 + \sum_{k=1}^K e^{\beta_0 + \beta_1 x_k}} \right)^2 \right\}.$$

For $K = 1$ and $\beta_0 = 0$, it gets simplified to

$$\frac{n}{(1 + e^{\beta_1 x_1})^2} \left[x_1^2 e^{\beta_1 x_1} \left(1 + e^{\beta_1 x_1}\right) - (x_1 e^{\beta_1 x_1})^2 \right] = n \frac{x_1^2 e^{\beta_1 x_1}}{(1 + e^{\beta_1 x_1})^2} \geq 0.$$

Therefore, we know that

$$-\frac{\partial^2 \ell(\beta_0, \beta_1)}{\partial \beta_1^2} \geq 0$$

for all β_0 and β_1, hence $-\ell(\beta_1)$ is convex.

As a result of the convexity of $-\ell(\beta_1)$, since the minimization is over $\beta_1 \in \mathbb{R}$ which is a convex space, GD with a suitable stepsize guarantees to find the MLE.

(You may use this blank page to continue your answer.)

5. (Bonus: 2 pts) Can you extend the result of the previous subquestion to $K \geq 2$?
Hint: for any two sequences $\{a_1, \ldots, a_n\}$ and $\{b_1, \ldots, b_n\}$, the Cauchy Schwarz inequality states that

$$\left(\sum_{i=1}^n a_i b_i\right)^2 \leq \left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right).$$

SOLUTION: For general $K \geq 2$, we have the claim by noting that

$$\begin{aligned}&\left(1 + \sum_{k=1}^K e^{\beta_0+\beta_1 x_k}\right) \sum_{k=1}^K x_k^2 e^{\beta_0+\beta_1 x_k}\\ &\geq \left(\sum_{k=1}^K e^{\beta_0+\beta_1 x_k}\right)\left(\sum_{k=1}^K x_k^2 e^{\beta_0+\beta_1 x_k}\right) \geq \left(\sum_{k=1}^K x_k e^{\beta_0+\beta_1 x_k}\right)^2.\end{aligned}$$

The last step uses the Cauchy Schwarz inequality.

For Sec0201, the solution keys are: AABDC CBAAD

Problem 5 (10 points, 1 point for each subquestion)

Be sure to mark your answers on the **Scantron Sheet** of multiple choice questions. There is **only one** correct answer to each question.

1. Which of the following statement is true

   A The Bayes classifier always has the smallest training error rate among all possible classifiers.

   B The Bayes classifier can not have zero training error rate.

   C The Bayes classifier always has the smallest false negative rate on the test data among all possible classifiers.

   D When Y has three categories (e.g., $Y \in \{1, 2, 3\}$), the Bayes error rate at any given feature point $X = x$ can not exceed $2/3$.

   SOLUTION: D

2. Which of the following statements about splines is not true?

   A A spline is a piecewise polynomial function of the original features.

   B Cubic splines with specified knots belong to the class of linear models (linearity in the parameters)

   C Splines can only be used when the original feature is one-dimensional.

   D Natural cubic splines has fewer parameters to estimate than cubic splines.

   SOLUTION: C

3. Which of the following statement is not true

   A Logistic regression is a parametric classification approach.

   B The maximum likelihood estimator (MLE) of the coefficients under logistic regression does not have an explicit expression.

   C The MLE under logistic regression can be computed by gradient descent with a suitable step size.

   D Logistic regression cannot be used when the response label has more than two classes.

   SOLUTION: D

4. Which of the following statements about gradient descent (GD) is true?

   A GD guarantees convergence to the global minimum for all types of objective functions.

   B GD is only applicable when the objective function is convex .

   C GD is only applicable when the objective function is differentiable.

   D In stochastic GD, the single data point used to compute the gradient must be distinct across iterations.

   SOLUTION: C

5. Which of the following statements is not true regarding Generalized Additive Models (GAMs)?

   A GAMs allow for the inclusion of non-linear functions of the original features.

   B GAMs can be used to model the interaction between the original features.

   C GAMs do not suffer from the curse of dimensionality.

   D GAMs can handle both continuous and categorical features.

   SOLUTION: B

6. Which of the following statement is not true

   A K-nearest neighbour (NN) is a nonparametric approach.

   B K-NN method has a large bias when K is small.

   C K-NN method does not require model-fitting.

   D K-NN method suffers from the curse of dimensionality.

   SOLUTION: B

7. Which of the following statements about the step size (learning rate) in gradient descent is true?

   A A larger step size always leads to faster convergence to the minimum.

   B A too small step size causes the algorithm to converge very slowly.

   C A too small step size leads to overfitting.

   D Using the training data to choose the step size leads to overfitting.

   SOLUTION: B

8. Which of the following approaches cannot be used to fit cubic splines

   A The ordinary least squares (OLS) estimation.

   B The ridge estimation.

   C The OLS followed by subset selection.

   D The maximum likelihood estimation.

   SOLUTION: D

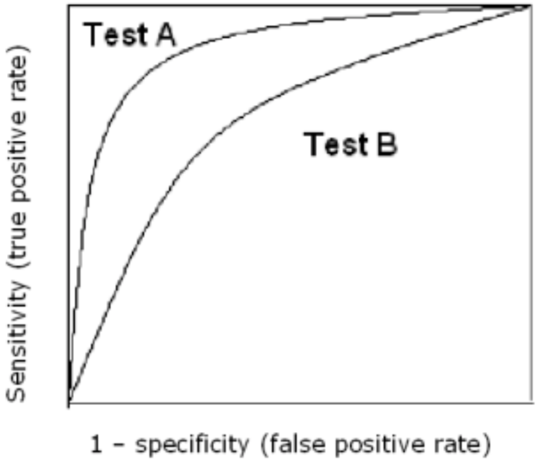
9. In the following plot, we compare two classification methods (called Test A and Test B) based on their ROC curves on the training data. Which of the following statement is true?

   A Test A has lower training error rate than Test B.

   B Test A has higher training error rate than Test B.

   C Test A has lower test error rate than B.

   D Test A is a more complex classifier than Test B.



SOLUTION: A

10. Which of the following statement is true between two classifiers

  A The classifier with lower false negative rate (FNR) must have higher false positive rate (FPR) comparing to the other.

  B The classifier with higher FNR must have larger misclassification error rate comparing to the other.

  C The classifier with both lower FNR and lower FPR does not necessarily have smaller misclassification error rate comparing to the other.

  D The classifier with lower FNR is less likely to misclassify positive data points as negative.

  SOLUTION: D

(You may use this page as scrach paper)

(You may use this page as scrach paper)

(You may use this page as scrach paper)

(You may use this page as scrach paper)