Statistical Machine Learning
Solutions for Exam 2020-08-22

1. i. **True**.
   ii. **True**.
   iii. **True**.
   iv. **True**.
   v. **False**. Random forest is an extension of bagging.
   vi. **False**. Gradient decent can converge to saddle points.
   vii. **True**.
   viii. **False**.
   ix. **False**.
   x. **True**.

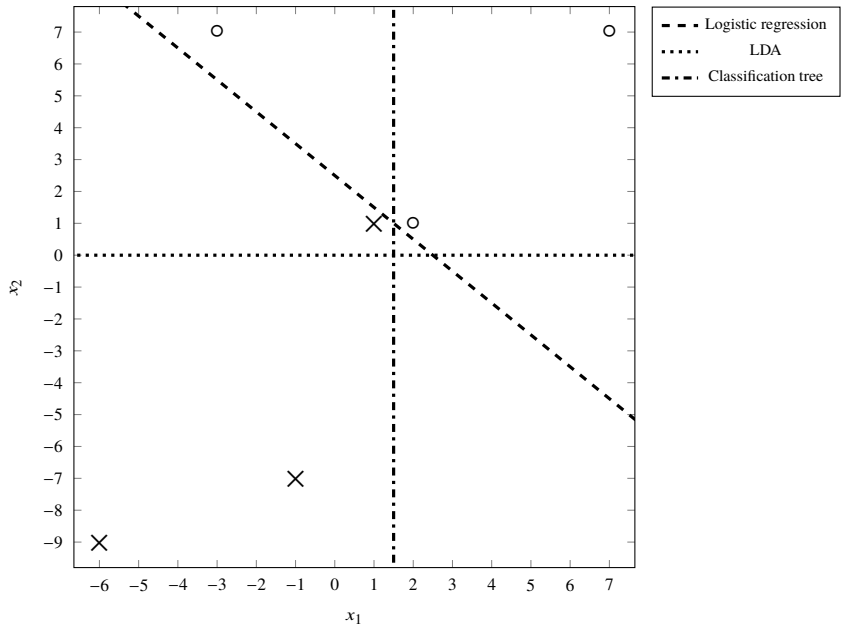

Figure 1: Visualization of the data and decision boundaries in problem 2.

2. (a) The data are visualized in Figure 1.

   (b) The decision boundary of a logistic regression classifier is given by $\boldsymbol{\theta}^\mathsf{T}\widetilde{\mathbf{x}} = 0$. We get

$$\boldsymbol{\theta}^\mathsf{T}\widetilde{\mathbf{x}} = 50 - 20x_1 - 20x_2 = 0 \iff x_2 = -x_1 + 2.5.$$

   The decision boundary is drawn in Figure 1 and then misclassification error is 0%.

   (c) Using the formula sheet, we get

$$\begin{aligned}
\widehat{\pi}_\times &= \frac{3}{6}, \\
\widehat{\pi}_\circ &= \frac{3}{6}, \\
\widehat{\boldsymbol{\mu}}_\times &= \frac{1}{3}\begin{bmatrix} -6-1+1 \\ -9-7+1 \end{bmatrix} = \begin{bmatrix} -2 \\ -5 \end{bmatrix}, \\
\widehat{\boldsymbol{\mu}}_\circ &= \frac{1}{3}\begin{bmatrix} -3+7+2 \\ 7+7+1 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}, \\
\widehat{\boldsymbol{\Sigma}}_{11} &= \frac{1}{6-2}\left((-4)^2 + 1^2 + 3^2 + (-5)^2 + 5^2 + 0^2\right) = 19 \\
\widehat{\boldsymbol{\Sigma}}_{12} &= \frac{1}{6-2}\left((-4)\cdot(-4) + 1\cdot(-2) + 3\cdot 6 + (-5)\cdot 2 + 5\cdot 2 + 0\cdot(-4)\right) = 8 \\
\widehat{\boldsymbol{\Sigma}}_{21} &= \widehat{\boldsymbol{\Sigma}}_{12}, \\
\widehat{\boldsymbol{\Sigma}}_{22} &= \frac{1}{6-2}\left((-4)^2 + (-2)^2 + 3^2 + 2^2 + 2^2 + (-4)^4\right) = 20,
\end{aligned}$$

From the hint, we know that decision boundary is perpendicular to the vector

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}^{-1}\left(\widehat{\boldsymbol{\mu}}_{\times}-\widehat{\boldsymbol{\mu}}_{\circ}\right) &= \begin{bmatrix}19 & 8\\ 8 & 20\end{bmatrix}^{-1}\begin{bmatrix}-4\\ -10\end{bmatrix}\\ &= \frac{1}{19\cdot 20-8\cdot 8}\begin{bmatrix}20 & -8\\ -8 & 19\end{bmatrix}\begin{bmatrix}-4\\ -10\end{bmatrix}\\ &= \frac{1}{381-64}\begin{bmatrix}-80+80\\ 32-190\end{bmatrix}\\ &= \frac{1}{317}\begin{bmatrix}0\\ -158\end{bmatrix}.\end{aligned}$$

This vector is vertical in the plane and thus the decision boundary must be horizontal. The decision boundary passes through the point

$$\frac{1}{2}\left(\begin{bmatrix}-2\\ -5\end{bmatrix}+\begin{bmatrix}2\\ 5\end{bmatrix}\right)=\begin{bmatrix}0\\ 0\end{bmatrix}.$$

Hence, the decision boundary is the horizontal line $x_2 = 0$ and it is shown in Figure 1. The classifier achieves $\frac{1}{6} = 17\%$ misclassification error.

(d) Since the problem states that we should only consider splits in the x_1 coordinate, we can directly see that any split between the $\times$ in (1,1) and the $\circ$in (2,1) achieves the optimal misclassification error of $\frac{1}{6} = 17\%$. The decision boundary for the split in $x_1 < 1.5$ is shown in Figure 1.

3. (a) Cost function (ii) does not penalize θ_1. This is consistent with B, where θ_1 does not tend to 0 as $\lambda \to \infty$.

   Cost function (i) uses regular ℓ_2-regularization (ridge regression) and (iii) uses ℓ_1-regularizatrion (LASSO). LASSO tends to give sparse solutions. That is, parameters become exactly equal to zero. We can see this in C, where, for example, θ_1 becomes 0 for $\lambda \gtrsim 100$.

   We get the final paring A-(1), B-(ii), C-(iii).

   (b) $\lambda \sum_{i=1}^{3} (\theta_i - \widetilde{\theta}_i)^2$, where $\widetilde{\boldsymbol{\theta}} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\mathsf{T}$.

   (c) For small values of λ, we have a small effect of the regularization and our model is more flexible. That means that we typically have a small error due to bias and a larger due to variance. As λ increases, large parameter values are penalized and our model becomes less flexible. This increases the error due to bias but decreases the error due to variance.

   When we have enough data, we can split the dataset prior to learning our model in a training and a test set. After learning the model using the training set, we can estimate the new data error on our test set. By relearning the model using different values of λ and evaluating using the test set, we can pick the λ that gives the smallest new data error.

4. (a) For linear regression, the learned parameter is given by $\widehat{\boldsymbol{\theta}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$. For model ii, we get

$$\mathbf{X} = \begin{bmatrix} \sin(\pi x_0) & \cos(\pi x_0) \\ \sin(\pi x_1) & \cos(\pi x_1) \\ \vdots & \vdots \\ \sin(\pi x_9) & \cos(\pi x_9) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \\ -1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & -1 \\ -1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\theta}}_{\text{ii}} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

   For model i, we get $\widehat{\theta}_{\text{i}} = 1$.

   (b) Since model i is a special case of model ii with $\theta_2 = 0$, model ii can always achieve the same training error as model i, but can also use θ_2 to improve the training error if possible.

   (c) By construction in this exercise, the off diagonal elements of $\mathbf{X}^\mathsf{T}\mathbf{X}$ will always be zero. Hence, $\widehat{\theta}_1$ will always be the same for both models.

   Split 1: Learn a model using $\{(x_i, y_i)\}_{i=0}^{4}$ and predict $\{(x_i, y_i)\}_{i=5}^{9}$. Let $\boldsymbol{\epsilon} = \mathbf{y} - \widehat{\mathbf{y}}$.

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ 7 \end{bmatrix} = \begin{bmatrix} 1 \\ 7/3 \end{bmatrix}. \\ \widehat{\mathbf{y}}_{\text{i}} &= \begin{bmatrix} 1 & 0 & -1 & 0 & 1 \end{bmatrix}^\mathsf{T} \\ \boldsymbol{\epsilon}_{\text{i}} &= \begin{bmatrix} 0 & 1 & 0 & -1 & 0 \end{bmatrix}^\mathsf{T} \\ \widehat{\mathbf{y}}_{\text{ii}} &= \begin{bmatrix} 1 & -7/3 & -1 & 7/3 & 1 \end{bmatrix}^\mathsf{T} \\ \boldsymbol{\epsilon}_{\text{ii}} &= \begin{bmatrix} 0 & 10/3 & 0 & -10/3 & 0 \end{bmatrix}^\mathsf{T} \\ \widehat{E}_{\text{new}}^{(\text{i})} &= 2 \\ \widehat{E}_{\text{new}}^{(\text{ii})} &= 200/9 \end{aligned}$$

   Split 2: Learn a model using $\{(x_i, y_i)\}_{i=5}^{9}$ and predict $\{(x_i, y_i)\}_{i=0}^{4}$.

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \\
\widehat{\mathbf{y}}_{\mathrm{i}} &= \begin{bmatrix} 0 & 1 & 0 & -1 & 0 \end{bmatrix}^{\mathsf{T}} \\
\boldsymbol{\epsilon}_{\mathrm{i}} &= \begin{bmatrix} 2 & 0 & -2 & 0 & 3 \end{bmatrix}^{\mathsf{T}} \\
\widehat{\mathbf{y}}_{\mathrm{ii}} &= \begin{bmatrix} -1 & 1 & 1 & -1 & -1 \end{bmatrix}^{\mathsf{T}} \\
\boldsymbol{\epsilon}_{\mathrm{ii}} &= \begin{bmatrix} 3 & 0 & -3 & 0 & 4 \end{bmatrix}^{\mathsf{T}} \\
\widehat{E}^{(\mathrm{i})}_{\mathrm{new}} &= 17 \\
\widehat{E}^{(\mathrm{ii})}_{\mathrm{new}} &= 34
\end{aligned}
$$

Summary:

$$
\begin{aligned}
\widehat{E}^{\mathrm{i}}_{\mathrm{new}} &= \frac{1}{2}\,(2 + 17) = 9.5 \\
\widehat{E}^{\mathrm{ii}}_{\mathrm{new}} &= \frac{1}{2}\,(200/9 + 34) = 253/9 \approx 28.1 > \widehat{E}^{\mathrm{i}}_{\mathrm{new}}
\end{aligned}
$$

Since the estimated new data error is larger for model ii, model i is the recommended model for in-production use.

5. (a) It straightforward to see that the given decision boundary in (a) is consistent with minimizing the misclassification error in each split.

   (b) By splitting the data at $x_1 < 2$ and then $x_2 > 2$, we correctly classify all training data, which is optimal for minimizing training misclassification error.

   (c) From the figure, we see that we have two possible splits along the x_1 and two along x_2.

	n_1	$\circ$	$\times$	Q_1	n_2	$\circ$	$\times$	Q_2	$n_1Q_1 + n_2Q_2$
$x_1 < 2$	6	3	3	$\frac{1}{2}$	7	0	7	0	3
$x_1 < 1$	3	2	1	$\frac{4}{9}$	10	1	9	$\frac{18}{100}$	$\frac{94}{30} > 3$
$x_2 < 2$	6	0	6	0	7	3	4	$\frac{24}{49}$	$\frac{24}{7} > 3$
$x_2 < 3$	9	1	8	$\frac{16}{81}$	4	2	2	$\frac{1}{2}$	$\frac{34}{9} > 3$

   The optimal first split is at $x_1 < 2$ For the second split, splitting at $x_2 < 2$ separates the training data perfectly and is the optimal second split.