Statistical Machine Learning
Solutions for Exam 2019-03-15

1. i. **True**
   ii. **False**, Regularization is used to reduce the variance and will thus increase the bias
   iii. **False**
   iv. **True**, It is linear in the parameters and thus a linear model
   v. **False**
   vi. **False**
   vii. **False**, Random forest is an bagging algorithm not a boosting algorithm
   viii. **False**
   ix. **False**
   x. **False**

2. (a) Linear regression, since that model is useful for predicting quantitative variables (i.e., a *regression* model)

   (b) **id** - Ignore since it carries no meaningful information

   bean - Input

   percentage - Input

   year - Input

   origin - Input

   producer - Input

   milk - Input

   timestamp - Ignore since it carries no meaningful information

   weight - Input

   price - Output since it is the thing we want to predict

   (c) **bean** - Qualitative

   percentage - Quantitative

   year - Could be viewed qualitative if the beans are of different qualities for different years or quantitative if the beans becomes bad with time

   origin - Qualitative

   producer - Qualitative, the numbers does have any orderly relationship

   milk - Qualitative, two options

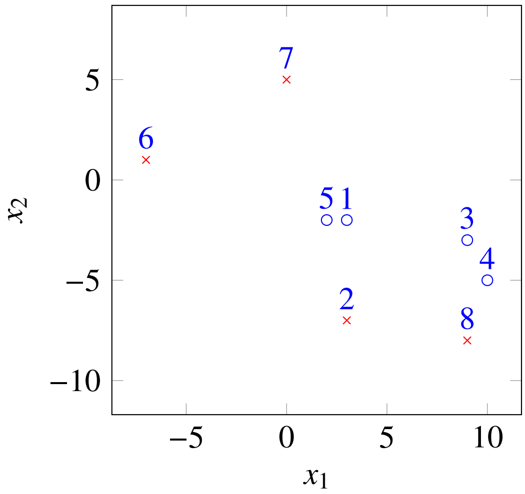
   weight - Quantitative

   price - Quantitative

   Note, these are only suggestions other correct answers may exist if well motivated

   (d) There are 395 different producers but we only have 183 examples meaning that there is a lot of producers we do not have examples for. We will therefore have a hard time predicting for new unseen producers

3. (a) The training data points are illustrated in the following graph.



(b) For $k = 1$ and $k = 3$ the estimated misclassification rate is 0.375 and 0.625, respectively. *(To score any points on this problem, you would also need to explain how these numbers were obtained.)*

(c) According to the estimated misclassification rates in (b), $k = 1$ is a better choice than $k = 3$ for this problem.

(d) See the course literature.

(e) If one uses an even k, the set of nearest neighbours can contain an equal amount of points from both classes. In this case it is not possible to determine the predicted class with a majority vote among these k nearest neighbours. This issue can, for instance, be handled by assigning one class at random in such a case.

4. (a) A linear classifier can have a nonlinear decision boundary in the *original input space* if the inputs are transformed using a nonlinear transformation. In this case a decision boundary corresponding to a circle centered in the origin would completely separate the two classes in Figure 1 and hence achieve zero misclassification error. This decision boundary can be achieved by using $x_1^2 + x_2^2$ as input instead of x_1 and x_2. The decision boundary for a linear classifier is always linear in its input space.

   (b) The decision boundary is given by the line at which we the model predicts $p(y = 0 \mid \mathbf{x}) = p(y = 1 \mid \mathbf{x})$, which for binary classification means $p(y = 1 \mid \mathbf{x}) = 0.5$ as the likelihood for either of the classes given x. I.e for either of the classes, $k \in \{0, 1\}$, we have

$$p(y = k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y = k)p(y = k)}{\sum_j p(\mathbf{x} \mid y = j)p(y = j)} = 0.5$$

   and thus

$$\frac{p(y = 0 \mid \mathbf{x})}{p(y = 1 \mid \mathbf{x})} = 1. \tag{1}$$

   Inserting the QDA normal assumption we get

$$\begin{aligned}\frac{p(y = 0 \mid \mathbf{x})}{p(y = 1 \mid \mathbf{x})} &= \frac{p(\mathbf{x} \mid y = 0)p(y = 0)}{p(\mathbf{x} \mid y = 1)p(y = 1)} \\ &= \frac{\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_0, \Sigma_0)p(y = 0)}{\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_1, \Sigma_1)p(y = 1)}\end{aligned}$$

   where $\boldsymbol{\mu}_*$ and Σ_* is the mean and variance for corresponding class. We will further denote $p(y = k)$ by π_k

   Inserting this into 1 and taking the logarithm of both sides we get

$$\log \frac{\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_0, \Sigma_0)\pi_0}{\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_1, \Sigma_1)\pi_1} = 0$$

$$\Longrightarrow$$

$$-\frac{1}{2}\log\det\Sigma_0 - \frac{1}{2}\|\mathbf{x} - \boldsymbol{\mu}_0\|^2_{\Sigma_0^{-1}} + \log\pi_0 + \frac{1}{2}\log\det\Sigma_1 + \frac{1}{2}\|\mathbf{x} - \boldsymbol{\mu}_1\|^2_{\Sigma_1^{-1}} - \log\pi_1 = 0$$

   We can now gather the terms depending on the degree $\mathbf{x}$ they contain by expanding the squares

$$\mathbf{x}^\mathbf{T}\underbrace{\frac{1}{2}(\Sigma_1^{-1} - \Sigma_0^{-1})}_{\text{B}}\mathbf{x} + \mathbf{x}^\mathbf{T}\underbrace{(\Sigma_0^{-1}\boldsymbol{\mu}_0 - \Sigma_1^{-1}\boldsymbol{\mu}_1)}_{v} +$$

$$\underbrace{\log\pi_0 - \log\pi_1 - \frac{1}{2}\log\det\Sigma_0 + \frac{1}{2}\log\det\Sigma_1 - \frac{1}{2}\boldsymbol{\mu}_0^\mathbf{T}\Sigma_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{2}\boldsymbol{\mu}_1^\mathbf{T}\Sigma_1^{-1}\boldsymbol{\mu}_1}_{-c} = 0$$

   and we are done.

   (c) See the course literature.

   (d) The inputs x_1 and x_2 could be correlated. In the extreme case, if x_2 is deterministically given by x_1 via the relationship $x_2 = (-0.9 + 4x_1)/5.1$, then the models (T1) and (T2) would be equivalent. Even if the dependence between x_1 and x_2 is not this extreme, we can still obtain a similar effect in the regression model.

5. (a) See the course literature.

(b) We want a classifier $\widehat{y}_{\text{boost}}(x)$ that changes from +1 to -1 somewhere between $x = 0$ and $x = 1$, say at $x = 0.5$, and back from -1 to $+1$ somewhere between $x = 1$ and $x = 2$, say at $x = 1.5$. For this we use the two base classifiers

$$\widehat{y}^1(x) = \text{sign}(x - 1.5),$$
$$\widehat{y}^2(x) = -\text{sign}(x - 0.5).$$

We choose $\alpha_1 = 1$ and $\alpha_2 = 2$ as confidence coefficients, other choices work equally well. With these two classifiers we have that

$$\sum_{b=1}^{2} \alpha_b \widehat{y}^b(x) = \begin{cases} -1 + 2 = 1, & x < 0.5 \\ -1 - 2 = -3, & 0.5 < x < 1.5 \\ 1 - 2 = -1, & x > 1.5 \end{cases} \tag{2}$$

To get a positive value also for $x > 1.5$ we add a third base classifier

$$\widehat{y}^3(x) = -\text{sign}(x - 3), \tag{3}$$

which will act as an offset. With $\alpha_3 = 2$ we get

$$\sum_{b=1}^{3} \alpha_b \widehat{y}^b(x) = \begin{cases} -1 + 2 + 2 = 3, & x < 0.5 \\ -1 - 2 + 2 = -1, & 0.5 < x < 1.5 \\ 1 - 2 + 2 = 1, & 1.5 < x < 3 \\ 1 - 2 - 2 = -3, & x > 3 \end{cases} \tag{4}$$

and all three training data points are correctly classified.

(c) Without loss of generality, assume that $\alpha_1 > \alpha_2$. The sign of the sum $\sum_{b=1}^{2} \alpha_b \widehat{y}^b(x)$ will then always equal to the sign of $\widehat{y}^1(x)$ regardless of the sign of $\widehat{y}^2(x)$. The classifier $\widehat{y}_{\text{boost}}(x)$ will then only have one switch in x (where $\widehat{y}^1(x)$ switches) and can consequently not completely separate the training data, for which at least two switches are needed.