Proef/oefen tentamen 2018, vragen en antwoorden

Machine Learning (Erasmus Universiteit Rotterdam)

1. (5 pts.) Given a data set, write down a pseudocode (the steps) that you will use to apply K-NN on a computer.

Solution:

0. Data loading.
1. Determine the value of **K**.
2. While reaching **K** number of training data points
   a. Calculate similarity between test point and training data points.
   b. Select the most similar one and store it.
3. Get the most frequent class of the selected training data points.
4. Return the predicted class.

2. (10 pts.) Derive the following equation.

$$E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \text{Var}\left[\hat{f}(x_0)\right] + \left[\text{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \text{Var}\left[\epsilon\right]$$

Solution:

$$\begin{aligned}
E\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] &= E\left[y_0^2 + \hat{f}^2(x_0) - 2y_0\hat{f}(x_0)\right] \\
&= E\left[y_0^2\right] + E\left[\hat{f}^2(x_0)\right] - E\left[2y_0\hat{f}(x_0)\right] \\
&= \text{Var}\left[y_0\right] + E\left[y_0\right]^2 + \text{Var}\left[\hat{f}(x_0)\right] + E\left[\hat{f}(x_0)\right]^2 - 2f(x)E\left[\hat{f}(x_0)\right] \\
&= \text{Var}\left[y_0\right] + \text{Var}\left[\hat{f}(x_0)\right] + \left((f(x))^2 - 2f(x)E\left[\hat{f}(x_0)\right] + E\left[\hat{f}(x_0)\right]^2\right) \\
&= \text{Var}\left[y_0\right] + \text{Var}\left[\hat{f}(x_0)\right] + \left(f(x) - E\left[\hat{f}(x_0)\right]\right)^2 \\
&= \text{Var}[\epsilon] + \text{Var}\left[\hat{f}(x_0)\right] + \left(\text{Bias}\left(\hat{f}(x_0)\right)\right)^2
\end{aligned}$$

where

$$\text{Var}[y] = E[(y - E[y])^2] = E[(y - f(x))^2] = E[(f(x) + \epsilon - f(x))^2] = E[\epsilon^2] = \text{Var}[\epsilon] + (E[\epsilon])^2 = \text{Var}[\epsilon]$$

3. (10 pts.) Consider the manufacturing cycle of a mobile phone resulting with a production batch of $1,000$ phones. Each process is followed by an inspection step. The main goal is to detect the defective phones correctly so that the customer satisfaction does not decrease. However, the process should not label either the non-defective ones as defective or the defective ones as non-defective. To detect the defective ones, a classifier (e.g. LDA) is trained with a certain probability threshold τ . According to the results, at the end of a cycle, the classifier labeled 120 phones as defective whereas 10% of these phones are misclassified. The rate is 5% for the non-defective 880 phones.

		Actual	
		True	False
Predicted	True	108	12
	False	44	836

a. Construct a confusion matrix.

b. Supposed that the manufacturer has decided to pay high attention to the customer satisfaction. How should τ be changed to modify inspection process?

> **Solution:** τ should be decreased in order to decrease number of FN (False Negative) samples.

c. How would the change in the τ result in the confusion matrix? In which cells you are expecting a decrease? In which cells you are expecting an increase? Indicate in the table below. Mark each cell with D (decrease) or I (Increase).

		Actual	
		True	False
Predicted	True	$\uparrow$	$\uparrow$
	False	$\downarrow$	$\downarrow$

4. (10 pts.) Derive the ridge regularization from the Bayesian point of view as we have discussed through one of the assignments.

> **Solution:** Assume that we are trying to find the β with the given data y. According to the Bayes Theorem, the posterior probability is calculated as
>
> $$P(\beta|y) = \frac{P(y|\beta)P(\beta)}{P(y)}.$$
>
> We are trying to maximize this posterior probability.
>
> $$\begin{aligned}
> \beta_{\text{MAP}} &= \arg\max_{\beta} P(\beta|y) && (1)\\
> &= \arg\max_{\beta} \frac{P(y|\beta)P(\beta)}{P(y)} && (2)\\
> &= \arg\max_{\beta} P(y|\beta)P(\beta) && (3)\\
> &= \arg\max_{\beta} \log(P(y|\beta)P(\beta)) && (4)\\
> &= \arg\max_{\beta} \log(P(y|\beta)) + \log(P(\beta)) && (5)
> \end{aligned}$$
>
> We can move from Equation (12) to Equation (13) since $P(y)$ is constant. For Ridge Regression, we select our prior distribution as Gaussian with variance μ^2 for each β_i. By inserting the Likelihood

function we can have the following:

$$\arg\max_{\beta} \log(P(y|\beta)) + \log(P(\beta)) \tag{6}$$

$$= \arg\max_{\beta} \left[\log \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_{i,1}+\cdots+\beta_p x_{i,p}))^2}{2\sigma^2}} + \log \prod_{j=0}^{p} \frac{1}{\mu\sqrt{2\pi}} e^{-\frac{\beta_j^2}{2\mu^2}} \right] \tag{7}$$

$$= \arg\max_{\beta} \left[-\sum_{i=1}^{n} \frac{(y_i-(\beta_0+\beta_1 x_{i,1}+\cdots+\beta_p x_{i,p}))^2}{2\sigma^2} - \sum_{j=0}^{p} \frac{\beta_j^2}{2\mu^2} \right] \tag{8}$$

$$= \arg\min_{\beta} \frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (y_i-(\beta_0+\beta_1 x_{i,1}+\cdots+\beta_p x_{i,p}))^2 + \frac{\sigma^2}{\mu^2} \sum_{j=0}^{p} \beta_j^2 \right] \tag{9}$$

$$= \arg\min_{\beta} \left[\sum_{i=1}^{n} (y_i-(\beta_0+\beta_1 x_{i,1}+\cdots+\beta_p x_{i,p}))^2 + \lambda^2 \sum_{j=0}^{p} \beta_j^2 \right] \tag{10}$$

where the likelihood function is defined as

$$L(\beta|y) := P(y|\beta) = \prod_{i=1}^{n} P_Y(y_i|\beta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_{i,1}+\cdots+\beta_p x_{i,p}))^2}{2\sigma^2}}.$$

5. (10 pts.) Compare AIC and BIC scores and specify when to use each one of them. Recall that

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2d\hat{\sigma}^2),$$
$$\text{BIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + \log(n)d\hat{\sigma}^2).$$

Solution: (3 of these gets full credit)

- BIC penalizes the model complexity more heavily.
- AIC can choose larger model since it puts less penalty on for an additional variable.
- To identify a good explanatory model, BIC should be used and to identify a good predictive model, AIC should be used.
- AIC aims to find a model that most accurately describes the data. On the contrary, BIC aims to discover the true model among the candidate sets.

6. a. (5 pts.) What are the two advantages of using AdaBoost as discussed in the paper entitled "*Aggregate features and AdaBoost for music classification?*"

Solution:

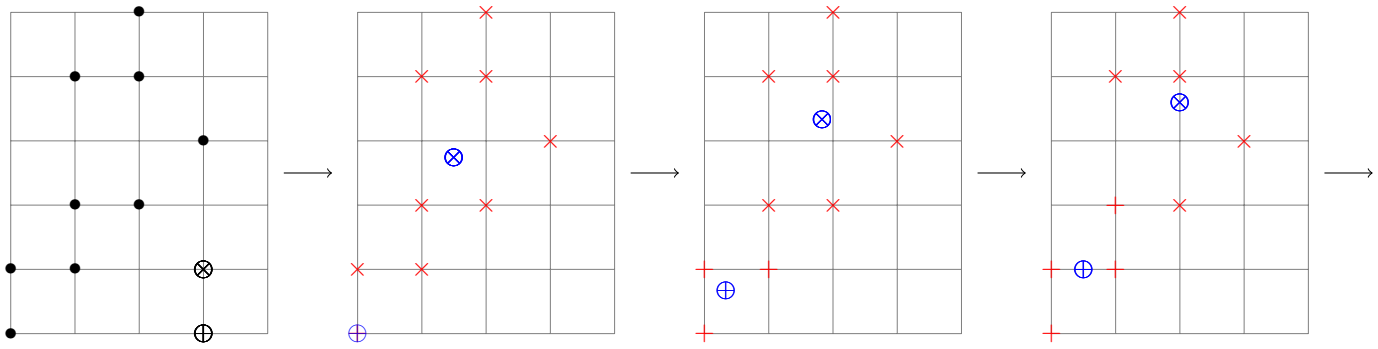
- Their model uses simple decision stumps operating on a single feature dimension. So, classification and feature selection are done in parallel.
- The model scales linearly with the number of training points provided when the iteration number is limited.

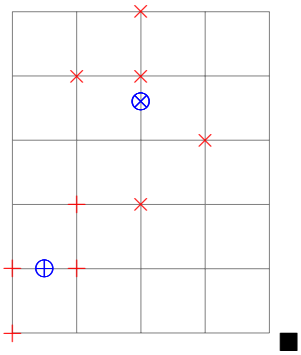
b. (10 pts.) Compare **k-fold** cross validation approach against the **leave-one-out** cross validation approach in terms of bias, variance and computation time.

Solution:

- k-fold CV
  - has computational advantage when $k < n$,
  - has less variance,
  - has higher bias,
  - gives more accurate estimates of the test error rate.
- leave-one-out CV
  - is computationally expensive,
  - has higher variance,
  - has almost no bias,
  - gives highly variable estimates.

7. (10 pts.) Group the points given in the grids below by using the **K-means** clustering method. You are supposed to have **two** clusters in the end. Describe the members of the clusters with $(+, \times)$, and the centroids of the clusters with $(\oplus, \otimes)$. Assume that the random initial centers of the clusters are given as $\oplus = (3, 0)$ and $\otimes = (3, 1)$.





8. Consider the likelihood function $\ell(\beta_0, \ldots, \beta_p)$ for logistic regression as we have discussed in the class.

   a. (4 pts.) Show that

$$\ell(\beta_0, \ldots, \beta_p) = \Pi_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}.$$

Solution: Assume that an event occurs with probability $p(x)$. Let y be an indicator variable and takes 1 if the event occurs. Since this event is a Bernoulli trial, we can rewrite the probability in the following form:

$$P[Y = y|X = x] = p(x)^y(1 - p(x))^{1-y}$$

. Therefore, for all data consisting n points, the likelihood function can be written as

$$\ell(\beta_0, \ldots, \beta_p) = \Pi_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i},$$

where i corresponds to the ith observation.

   b. (6 pts.) Using part a. above show that

$$\frac{\partial \log(\ell(\beta_0, \ldots, \beta_p))}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - p(x_i))x_{ij}.$$

Solution: From part a, we have

$$\ell(\beta_0, \ldots, \beta_p) = \Pi_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}.$$

If we take the logarithm of the function, the result would be

$$\begin{aligned}
\log\left(\ell(\beta_0, \ldots, \beta_p)\right) &= \log\left(\Pi_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}\right) \\
&= \sum_{i=1}^{n} y_i \log p(x_i) + (1 - y_i)\log(1 - p(x_i)) \\
&= \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1 - p(x_i)} \\
&= \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i(\beta_0 + \beta^T x_i) \\
&= \sum_{i=1}^{n} -\log 1 + e^{\beta_0 + \beta^T x_i} + \sum_{i=1}^{n} y_i(\beta_0 + \beta^T x_i).
\end{aligned}$$

One would achieve the following equivalence with taking the partial derivative of log likelihood function with respect to the component β_j.

$$\begin{aligned}\frac{\partial \log(\ell(\beta_0, \ldots, \beta_p))}{\partial \beta_j} &= -\sum_{i=1}^{n} \frac{1}{1 + e^{\beta_0 + \beta^T x_i}} e^{\beta_0 + \beta^T x_i} x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \\ &= \sum_{i=1}^{n} (y_i - p(x_i)) x_{ij}.\end{aligned}$$

9. Suppose that you are given the data set in Table 2. Then answer the following classification tree questions.

	x_1	x_2	class
S_1	black	0	1
S_2	black	1	1
S_3	red	0	2
S_4	black	1	1
S_5	red	0	1
S_6	red	1	2
S_7	red	0	2
S_8	black	1	1
S_9	red	1	1
S_{10}	red	0	1

Table 1: Dataset

a. (2 pts.) Specify the splits in your tree.

Solution:
Splits on x_1: black/red.
Splits on x_2: 0/1.

b. (4 pts.) Calculate explicitly the gain for each split according to **Gini** index. Which split should be selected?

Solution: Formulation of the Gini index at category m is given as $\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$. You need to calculate weighted average of each category connected to the node to find the Gini value of a split.

$$\begin{aligned}I(x_1 = \text{red}) &= \frac{6}{10}\left[\left(\frac{3}{6} \times \frac{3}{6}\right) + \left(\frac{3}{6} \times \frac{3}{6}\right)\right] + \frac{4}{10}(1 \times 0)(0 \times 1) \\ &= 0.30 \\ I(x_2 = 0) &= \frac{5}{10}\left[\left(\frac{3}{5} \times \frac{2}{5}\right) + \left(\frac{2}{5} \times \frac{3}{5}\right)\right] + \frac{5}{10}\left[\left(\frac{4}{5} \times \frac{1}{5}\right) + \left(\frac{1}{5} \times \frac{4}{5}\right)\right] \\ &= 0.40\end{aligned}$$

Then, select $I(x_1 = \text{red})$ as a split since it has the minimum Gini index.

c. (4 pts.) Calculate explicitly the gain for each split according to **Entropy**. Which split should be selected?

Solution: Formulation of the Entropy index at category m is given as $-\sum_{k=1}^{K} \hat{p}_{mk}(\log \hat{p}_{mk})$. You need to calculate weighted average of each category connected to the node to find the Entropy value of a split.

$$\begin{aligned} I(x_1 = \text{red}) &= P_{\text{Yes}} I(\text{Yes}) + P_{\text{No}} I(\text{No}) \\ &= \frac{6}{10}\left(-\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}\right) + \frac{4}{10}\left(-\frac{4}{4}\log\frac{4}{4} - \frac{0}{4}\log\frac{0}{4}\right) \\ &= 0.42 \\ I(x_2 = 0) &= P_0 I(0) + P_1 I(1) \\ &= \frac{5}{10}\left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right) + \frac{5}{10}\left(-\frac{4}{5}\log\frac{4}{5} - \frac{1}{5}\log\frac{1}{5}\right) \\ &= 0.59 \end{aligned}$$

Then, select $I(x_1 = \text{red})$ as a split since it has the minimum entropy.

10. For each of parts a. through d. below, indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

    a. The sample size n is extremely large, and the number of predictors p is small.

    Solution: We may want to investigate some nonlinear relations. Therefore flexible methods are preferred. Overfitting is less of a concern.

    b. The number of predictors p is extremely large, and the number of observations n is small.

    Solution: An inflexible method (or a flexible method with regularization) would be a better fit.

    c. The relationship between the predictors and response is highly nonlinear.

    Solution: Since inflexible methods generally do not include nonlinear relations, the flexible methods are preferred.

    d. The variance of the error terms, *i.e.* $\text{Var}(\epsilon)$, is extremely high.

    Solution: With inflexible models, we cannot ensure to contain the high variance. It is better to use a flexible method.

11. Write down the complete optimality conditions (KKT) of the optimization problem on slide 13 of the lecture notes 6 (for $d = 2$).

Solution:

$$\begin{aligned} \min \quad & \frac{1}{2}\beta^T\beta + c\sum_{i=1}^{n}\varepsilon_i^2 \\ \text{s.t.} \quad & \\ & y_i(\beta^T x_i + \beta_0) \geq 1 - \varepsilon_i, && i = 1, 2, \ldots, n \\ & \varepsilon_i \geq 0 && i = 1, 2, \ldots, n \end{aligned}$$

The Lagrangian function would be

$$\mathcal{L}(\beta, \beta_0, \varepsilon; \alpha, \upsilon) = \frac{1}{2}\beta^T\beta + c\sum_{i=1}^{n}\varepsilon_i^2 - \sum_{i=1}^{n}\alpha_i\left(y_i(\beta^T x_i + \beta_0) - 1 + \varepsilon_i\right) - \sum_{i=1}^{n}\upsilon_i\varepsilon_i$$

where α_i and υ_i are the Lagrange multipliers.

First order conditions for optimality (except complementary slackness and feasibility):

$$\frac{\partial\mathcal{L}}{\partial\beta} = 0 \Rightarrow \qquad \beta = \sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial\mathcal{L}}{\partial\beta_0} = 0 \Rightarrow \qquad \beta_0 = \sum_{i=1}^{n}\alpha_i y_i = 0$$

$$\frac{\partial\mathcal{L}}{\partial\varepsilon_i} = 0 \Rightarrow \qquad \varepsilon_i = \frac{\alpha_i + \upsilon_i}{2c}, \qquad i = 1, 2, \ldots, n.$$

12. Suppose that you are going to train a two-layer feed-forward neural network for classifying a set of points on a plane. The points are labeled as black and white. Draw the network structure and specify the activation functions (you are free to select any number of hidden nodes).

Solution:

We have chosen to use two nodes in the hidden layer.