COURSE 02457

Non-Linear Signal Processing: Quiz Exercise 6

C.M. Bishop: Pattern Recognition and Machine Learning, Sections 1.5, 4.2, 4.3.4, 5.1-5.4

Questions

1. Consider the probability of class 1, C_1, given a data point, x, $P(C_1|x)$ in a binary classification setting. How would the probability $P(C_1|x)$ change if the probability of the data point given class 2, $P(x|C_2)$, were to increase? Assume all other quantities $P(C_1)$, $P(C_2)$, and $P(x|C_1)$ remain the same.

   (a) $P(C_1|x)$ would not be affected.

   (b) $P(C_1|x)$ would increase.

   (c) $P(C_1|x)$ would decrease

   (d) $P(C_1|x)$ would decrease, but so would $P(C_2|x)$ so that the ratio between the two would remain the same.

2. Do the error functions of logistic regression and multinomial regression have global, unique minima?

   (a) logistic regression: yes, multinomial regression: yes

   (b) logistic regression: yes, multinomial regression: no

   (c) logistic regression: no, multinomial regression: yes

   (d) logistic regression: no, multinomial regression: no

3. Which assumptions on class distributions lead to linear decision boundaries in classification problems?

   (a) Class distributions are Gaussian

   (b) Class distributions have the same covariance matrix

   (c) Class distributions have different covariance matrices

   (d) Class distributions have the same covariance matrix and are Gaussian

4. The standard form for linear equations with n variables is $a_1x_1 + a_2x_2 + \ldots + a_nx_n = b$, where $a_1, a_2, \ldots, a_n$, and b are constants. Let $\mathbf{a} = (a_1, a_2, \ldots, a_n)^T$ be the column vector containing the coefficients $a_1, a_2, \ldots, a_n$ of the variables. Then the standard form can be written $\mathbf{a}^T\mathbf{x} = b$, where $\mathbf{x}$ is the column vector containing the variables.

   Consider equation (4.65) in Bishop, which states $P(C_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0)$. This can be re-arranged to make it clear that it is a linear equation when $P(C_1|\mathbf{x})$ is constant, as is the case on decision boundaries. Re-arranging and identifying the resulting terms with those in the standard form for linear equations gives

(a) $\mathbf{w}$ corresponds to $\mathbf{a}$ in the standard form and $w_0 + \sigma^{-1}(P(C_1|\mathbf{x}))$ to b

(b) $\mathbf{w}$ corresponds to $\mathbf{a}$ in the standard form and $\sigma^{-1}(P(C_1|\mathbf{x})) - w_0$ to b

(c) $\mathbf{x}$ corresponds to $\mathbf{a}$ in the standard form and $w_0 + \sigma^{-1}(P(C_1|\mathbf{x}))$ to b

(d) $\mathbf{x}$ corresponds to $\mathbf{a}$ in the standard form and $\sigma^{-1}(P(C_1|\mathbf{x})) - w_0$ to b

5. Consider the cancer treatment loss function in Figure 1.25 in Bishop. The hospital has a probabilistic classifier $p(C_k|\mathbf{x})$ which takes the biomarker measurement $\mathbf{x}$ and predicts probability for C_1 = cancer and C_2 = normal. For a given patient the classifier gives $p(\text{cancer}|\mathbf{x}) = 1/1000$. What should we do?

   (a) Treat the patient as if the patient is normal because that has the lowest associated expected loss.

   (b) Treat the patient as if the patient has cancer because that has the lowest associated expected loss.

   (c) The expected losses of normal and cancer are so close that we cannot make a decision.

   (d) We cannot use the classifier to make decisions because its predictions are not 100 % certain.

6. In a neural network for binary classification the output

$$y(\mathbf{x}|\mathbf{w}) = \frac{1}{1 + \exp(-a(\mathbf{x};\mathbf{w}))} = \sigma(\, a(\mathbf{x};\mathbf{w})\,)$$

   gives the probability for class 1. What is $a(\mathbf{x};\mathbf{w})$?

   (a) $\mathbf{w} \cdot \mathbf{x}$

   (b) $\sigma(\, \mathbf{w}^{(2)} \cdot \mathbf{z}\,)$

   (c) $\tanh(\, \mathbf{w}^{(1)} \cdot \mathbf{x}\,)$

   (d) $\mathbf{w}^{(2)} \cdot \mathbf{z}$,

   where $\mathbf{z}$ is shorthand for the output of the hidden unit.

7. Which of the following statements are correct?

   I) A discriminative function can be used to simulate data.

   II) Posterior probabilities of class membership are found both when using discriminative models and generative models.

   III) Discriminant functions map explanatory variables directly onto class labels without using probabilities.

   (a) I)

   (b) I) and II)

   (c) II) and III)

   (d) I), II), and III)

8. When using the squared loss as cost function in regression problems, what is the best prediction?

   (a) The conditional mean, conditioning on target variables used during training

   (b) The conditional mean, conditioning on explanatory variables used during training

   (c) The conditional mean, conditioning on the observed explanatory variables for which a prediction is desired

   (d) The conditional mean, conditioning on the observed explanatory variables for which a prediction is desired, plus the intrinsic variance of the data

9. When using the squared loss as cost function in regression problems, what is the minimal attainable error?

   (a) Zero

   (b) The precision of the noise on the target variables

   (c) The standard deviation of the noise on the target variables

   (d) The variance of the noise on the target variables

10. In multi-class logistic regression, we have that $P(C_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$, where $a_k = \mathbf{w}_k^T\phi$, where $\mathbf{w}_k$ is the vector of coefficients and ϕ the vector of basis functions of explanatory variables. What is $\frac{\partial y_k}{\partial a_j}$?

    (a) $y_k(1-y_k)$ when $k=j$ and $-y_k y_j$ when $k \neq j$

    (b) $\exp(a_k)\phi$ when $k=j$ and $\exp(a_j)\phi$ when $k \neq j$

    (c) $\exp(a_k)\mathbf{w}$ when $k=j$ and $\exp(a_j)\mathbf{w}$ when $k \neq j$

    (d) $y_k^2(1-y_k)$ when $k=j$ and $-y_k^2 y_j$ when $k \neq j$

Hint - reading for each question

1. Section 4.2
2. Section 4.3.4
3. Section 4.2.1
4. Section 4.2.1
5. Section 1.5.2
6. Section 5.1
7. Section 1.5.4
8. Section 1.5.5
9. Section 1.5.5
10. Section 4.3.4

DTU, October 2013,

Laura Frølich and Ole Winther