DATA 1010
In-class exercises
Samuel S. Watson
09 November 2018

Problem 1

I did some consulting once for a businessman who told me that the way to distinguish a machine learning expert from a novice is to give them some of your data and see whether they can write a classifier which has very low error when applied to a withheld test set. Why was he wrong?

Problem 2

Consider a binary classification problem where the two classes are equally probable, the class-0 conditional density is a standard multivariable normal distribution in two dimensions, and the class-1 conditional density is a multivariate normal distribution with mean $[1, 1]$ and covariance I.

Find the class boundary for the Bayes classifier.

Problem 3

(Problem 2 continued). Find the regression function $r(x) = \mathbb{E}[Y \mid X = x] = \mathbb{P}(Y = 1 \mid X = x)$. Plot a heatmap of this function.

Problem 4

(Problem 2 continued). Sample 1000 points by choosing one of the two distributions uniformly at random and then sampling from the selected distribution. Model the regression function by fitting the parametric model

$$r(\mathbf{x}) = \sigma(\alpha + \boldsymbol{\beta} \cdot \mathbf{x}),$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$. Use the loss functional

$$L(r) = \sum_{i=1}^{n} \left[y_i \log \frac{1}{r(x_i)} + (1 - y_i) \log \frac{1}{1 - r(x_i)} \right].$$

This is called **logistic regression**.

Problem 5

(Problem 2 continued). Was the supposition that r took the given parametric form correct?

Problem 6

How could we have modified the setup of Problem 2 so that our parametric assumption did not hold? Hint: begin by showing that the decision boundary is necessarily linear for a logistic model.

Problem 7

How could we modify the logistic regression model so that it would accommodate multivariate normal class conditional distributions with distinct covariance matrices?