Learning from Data, Tutorial Sheet 3

School of Informatics, University of Edinburgh

Instructor: Amos Storkey

The Gaussian distribution in one dimension is defined as

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \text{ , where } \int_{-\infty}^{\infty} p(x)dx = 1$$

1. We know that the mean is defined as $\int_{-\infty}^{\infty} xp(x)dx = \mu$ and the variance is $\int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx = \sigma^2$. Show that the distribution of variable $y = (x-\mu)/\sigma$ has zero mean and unit variance. Comment on the normalisation constant and ensure you are familiar with the rules for change of variables for densities (using the Jacobian).

2. Consider data $x^i, i = 1, \ldots, P$. Given mean μ, show that the Maximum Likelihood estimator of σ^2 is $\hat{\sigma^2} = \frac{1}{P}\sum_{i=1}^{P}(x^i - \mu)^2$

3. A training set consists of one dimensional examples from two classes. The training examples from class 1 are $\{0.5, 0.1, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.35, 0.25\}$ and from class 2 are $\{0.9, 0.8, 0.75, 1.0\}$. Fit a (one dimensional) Gaussian using Maximum Likelihood to each of these two classes. Also estimate the class probabilities p_1 and p_2 using Maximum Likelihood. What is the probability that the test point $x = 0.6$ belongs to class 1?

4. Load the data in the file `week4.mat` into matlab. The variable `x1` gives the attributes for each data point belonging to class 1. Each column is an attribute and each row a data point. The variable `x2` gives the data for class 2. Write some matlab code to train a class conditional Gaussian classifier on `x1` and `x2`. Use this to classify the data in `xtest`. Plot a class conditional colour plot of `xtest` for the first two attributes and describe the decision boundary.

5 (Harder) Given the distributions $p(x|class1) = N(\mu_1, \sigma_1^2)$ and $p(x|class2) = N(\mu_2, \sigma_2^2)$, with corresponding prior occurrence of classes p_1 and p_2 ($p_1 + p_2 = 1$), calculate the decision boundary explicitly as a function of $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p_1, p_2$. How many solutions are there to the decision boundary, and are they all reasonable?