Appendix A

Mathematical stuff

A.1 Differentiation and vectors

Suppose we have a vector $\mathbf{w}$ and a function $f(\mathbf{w})$. Differentiation of f with respect to $\mathbf{w}$ is defined as,

$$\frac{\partial}{\partial \mathbf{w}} f(\mathbf{w}) = \frac{\partial f}{\partial \mathbf{w}} \equiv \left(\frac{\partial f}{\partial \omega_1}, \frac{\partial f}{\partial \omega_2}, \ldots, \frac{\partial f}{\partial \omega_N} \right)^T$$

Example 1:

$$\begin{aligned} f(\mathbf{w}) &= \mathbf{x}^T \mathbf{w} \qquad \text{(inner product)} && \text{(A.1)} \\ f(\mathbf{w}) &= \sum_{i=1}^{N} x_i \omega_i \qquad \Rightarrow && \text{(A.2)} \\ \frac{\partial f}{\partial \omega_i} &= x_i \qquad \Rightarrow \qquad \frac{\partial f}{\partial \mathbf{w}} = \mathbf{x} && \text{(A.3)} \end{aligned}$$

Example 2:

$$\begin{aligned} f(\mathbf{w}) &= \mathbf{w}^T \mathbf{R} \mathbf{w} \qquad (\mathbf{R} \text{ is a symmetric } N \text{ by } N \text{ matrix}) && \text{(A.4)} \\ f(\mathbf{w}) &= \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_i R_{ij} \omega_j \qquad \Rightarrow && \text{(A.5)} \\ \frac{\partial f}{\partial \omega_i} &= 2 \sum_{j=1}^{N} R_{ij} \omega_j \qquad \Rightarrow && \text{(A.6)} \\ \frac{\partial f}{\partial \mathbf{w}} &= 2 \mathbf{R} \mathbf{w} && \text{(A.7)} \end{aligned}$$

Appendix B

Statistics

B.1 Bays Theorem

This section is intended as a small introduction to Bays probabilities and Bays theorem. We will use a small example for illustration [1].

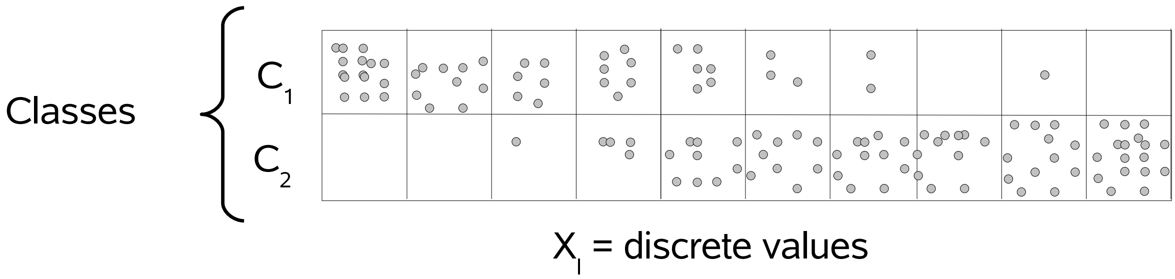

Figure B.1: Data from a set of images and the measurements. The measurement for a given image can result in one of a discrete set of values $\{X_l\}$ and the images can belong to one of two possible classes C_1 and C_2. The number of dots in each cell represents the number of images with the values X_l and corresponding class label. Various probabilities can be computed by the fractions of points falling in different regions of the cells.

We have images that belongs to two different classes, C_1 and C_2. A measure of the image results in one of a discrete set of values $\{X_l\}$. The set of images and their measurements is shown in Fig. B.1.

GOAL: To classify a new images in such a way as to minimize the probability of misclassi-

[1]This example is taken from the book of Bishop (*Neural Networks for Pattern Recognition*)

fication.

Introduce

$$P(C_k) = \text{ prior probability of an image belonging to class } C_k$$

If we dont have any measurements of the new image we can use the prior probabilites and assign the new images to class C_1 if $P(C_1) > P(C_2)$ otherwise class C_2. How do you compute $P(C_1)$ and $P(C_2)$ using Fig. B.1? Now suppose we have a measurement then introduce,

$$\begin{aligned} P(C_k, X_l) &= \text{ joint probability, i.e. probability that an image} \\ &\quad\ \text{belongs to } C_l \text{ AND has the measurement } X_l. \end{aligned}$$

$$\begin{aligned} P(X_l|C_k) &= \text{ conditional probability, i.e. the probability of measurement} \\ &\quad\ X_l \text{ GIVEN that the image belongs to class } C_k. \end{aligned}$$

It is quite obvious that

$$\begin{aligned} P(C_k, X_l) &= P(X_l|C_k)\ P(C_k) \\ P(C_k, X_l) &= P(C_k|X_l)\ P(X_l) \end{aligned}$$

One can also understand these relations by counting points in the different cells in Fig. B.1. Combining these two expression results in the famous Bays's theorem,

$$P(C_k|X_l) = \frac{P(X_l|C_k)P(C_k)}{P(X_l)} \tag{B.1}$$

where $P(C_k|X_l)$ is called the posterior probability and $P(X_l|C_k)$ is called the class-conditional probability. Usually we are interested in $P(C_k|X_l)$ since it gives us the probability of class C_k given the measurement X_l.

Note that any new mesurement X_l must be assigned to one of the two classes,

$$P(C_1|X_l) + P(C_2|X_l) = 1$$

Using this together with Eqn. B.1 we obtain

$$P(X_l) = P(X_l|C_1)P(C_1) + P(X_l|C_2)P(C_2) \quad \Big(= P(X_l, C_1) + P(X_l, C_2)\Big)$$

which again is quite obvious.

A word about continuous variables. Most of the time the measurements are continuous which means that we have a probability density function $p(x)$. The probability for x being in an interval is as usual,

$$P(x \in [a, b]) = \int_a^b p(x)dx$$

The class-conditional probability becomes the class-conditional probability density.

$$P(C_k|x) = \frac{p(x|C_k)P(C_k)}{p(x)}$$