Lecture Outline

Reading: Chapter 4 - Linear Models for Classification

This lecture covers

- Introduction
- 2-class discriminant functions
- K-class discriminant functions

4 Introduction to Linear Models for Classification

The goal in classification is to take an input vector $\mathbf{x}$ and to assign it to one of K discrete classes $\mathcal{C}_k$ where $k = 1, \ldots, K$.

Figure 1: Basic classifier

In the most common scenario, the classes are taken to be disjoint, so that each input is assigned to one and only one class. The input space is thereby divided into *decision regions* whose boundaries are called *decision boundaries* or *decision surfaces*.

Figure 2: Data space, decision regions and decision boundaries

In this chapter, we consider linear models for classification, by which we mean that the decision surfaces are linear functions of the input vector $\mathbf{x}$ and hence are defined by $(D-1)$-dimensional hyperplanes within the D-dimensional input space. Data sets whose classes can be separated exactly by linear decision surfaces are said to be *linearly separable*.

There are various ways of using target values to represent class labels. For probabilistic models, the most convenient, in the case of two-class problems, is the binary representation in which there is a single target variable $t \in \{0, 1\}$ such that $t = 1$ represents class $\mathcal{C}_1$ and $t = 0$ represents class $\mathcal{C}_2$.

For $K > 2$ classes, it is convenient to use a 1-of-K coding or "one hot" encoding scheme in which $\mathbf{t}$ is a vector of length K such that if the class is $\mathcal{C}_j$, then all elements t_k of $\mathbf{t}$ are zero except element t_j, which

Figure 3: Target for 2-class, 5-class problem

takes the value 1, i.e. $t_k = \delta(j-k)$. For instance if we have $K = 5$ classes, then a pattern (data point) from class 2 would be given the target vector

$$\mathbf{t} \quad = \quad (0, 1, 0, 0, 0)^T. \tag{1}$$

In Section 1.5.4, we identified three distinct approaches to the classification problem in terms of increasing order of complexity:

1. **Construct a discriminant function** Find a *discriminant function*, $f(\mathbf{x})$ that directly assigns each vector $\mathbf{x}$ to a specific class (Section 4.1).
2. **Discriminative approach** Solve the inference problem by directly modeling the posterior class probabilities, $p(\mathcal{C}_k|\mathbf{x})$ and then use these to make optimal classification decisions.
3. **Generative approach** First solve the inference problem of determining the class-conditional densities, $p(\mathbf{x}|\mathcal{C}_k)$ for each class $\mathcal{C}_k$ individually. Also separately infer the prior class probabilities $p(\mathcal{C}_k)$. The use Bayes' theorem

$$p(\mathcal{C}_k|\mathbf{x}) \quad = \quad \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \tag{2}$$

to find the posterior class probabilities $p(\mathcal{C}_k|\mathbf{x})$ to make optimal classification decisions. Approaches the explicitly or implicitly model the distribution of inputs as well as outputs are known as *generative models*, because by sampling from them it is possible to generate synthetic data points in the input space.

In the linear regression models considered in Chapter 3, the model prediction $y(\mathbf{x}, \mathbf{w})$ was given by a linear function of the parameters $\mathbf{w}$. In the simplest case, the model is also linear in the input variables and therefore takes the form $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$, so that y is a real number. For classification problems, however, we wish to predict *discrete class labels*, or more generally posterior probabilities that lie in the range $(0, 1)$. To achieve this, we consider a generalization of this model in which we transform the linear function of $\mathbf{w}$ using a nonlinear function $f(\cdot)$ so that

$$y(\mathbf{x}) \quad = \quad f(\mathbf{w}^T\mathbf{x} + w_0). \tag{3}$$

In the machine learning literature, $f(\cdot)$ is known as an *activation function*, whereas its inverse is called a *link function* in the statistics literature. Note, that in contrast to the models used for regression, they are no longer linear in the parameters due to the presence of the nonlinear function $f(\cdot)$ but the decision surfaces are linear.

The algorithms discussed in this chapter are equally applicable if we first make a fixed nonlinear transformation of the input variables using a vector of basis functions $\phi(\mathbf{x})$ as we did for the regression models in Chapter 3, i.e.

$$y(\mathbf{x}) \quad = \quad f(\mathbf{w}^T\phi(\mathbf{x}) + w_0). \tag{4}$$

Figure 4: Turning real-valued labels into discrete classes or posterior probabilities in $(0, 1)$

4.1 Discriminant Functions

A *discriminant* is a function that takes an input vector, $\mathbf{x}$ and assigns it to one of K classes, denoted $\mathcal{C}_k$. We shall restrict ourselves to *linear discriminants*, namely those for which the decision surfaces are hyperplanes.

4.1.1 Two classes

The simplest linear discriminant function is

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \tag{5}$$

where $\mathbf{w}$ is called a *weight vector* and w_0 is a bias. An input vector $\mathbf{x}$ is classified as follows:

$$\begin{cases} \mathcal{C}_1, & y(\mathbf{x}) \geq 0 \\ \mathcal{C}_2, & y(\mathbf{x}) < 0 \end{cases} \tag{6}$$

The corresponding decision surface is therefore defined by the relation $y(\mathbf{x}) = 0$ which corresponds to a $(D-1)$-dimensional hyperplane within the D-dimensional input space.

Figure 5: Decision surface

Weight vector, w determines orientation of the decision surface: Consider two different points $\mathbf{x}_A$ and $\mathbf{x}_B$ both of which lie on the decision surface. Because $y(\mathbf{x}_A) = y(\mathbf{x}_B) = 0$, we have

$$\begin{aligned} 0 &= y(\mathbf{x}_A) - y(\mathbf{x}_B) \\ &= \mathbf{w}^T\mathbf{x}_A + w_0 - \mathbf{w}^T\mathbf{x}_B - w_0 \\ &= \mathbf{w}^T(\mathbf{x}_A - \mathbf{x}_B) \end{aligned} \tag{7}$$

and hence the vector $\mathbf{w}$ is orthogonal to every vector lying within the decision surface, i.e. $\mathbf{w}$ is orthogonal to the decision surface. Therefore $\mathbf{w}$ determines the orientation of the decision surface (see Fig. 6).

Bias, w_0 determines location of the decision surface: If $\mathbf{x}$ is a point on the decision surface, then $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0$ or $\mathbf{w}^T\mathbf{x} = -w_0$ and thus the normal distance from the origin to the decision surface is given by

$$\frac{\mathbf{w}^T\mathbf{x}}{\|\mathbf{w}\|} = -\frac{w_0}{\|\mathbf{w}\|}. \tag{8}$$

Therefore the bias parameter w_0 determines the location of the decision surface. See Fig. 6 for an illustration.

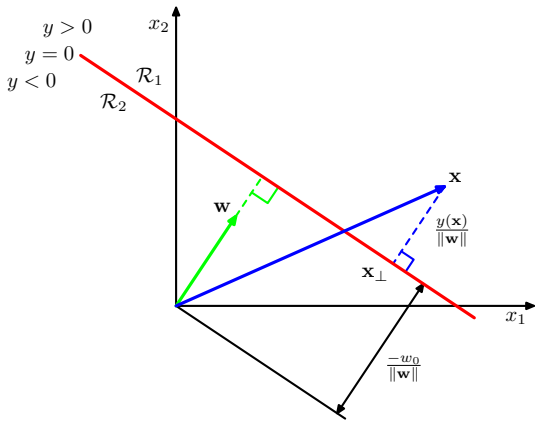

Figure 6:

$r = y(\mathbf{x})/\|\mathbf{w}\|$ **gives the signed distance of x from the decision surface**: Consider an arbitrary point $\mathbf{x}$ and let $\mathbf{x}_\perp$ be its orthogonal projection onto the decision surface, so that (see Fig. 6)

$$\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{\|\mathbf{w}\|}. \tag{9}$$

We then have

$$\begin{aligned} \mathbf{w}^T\mathbf{x} + w_0 &= \mathbf{w}^T\mathbf{x}_\perp + w_0 + r\mathbf{w}^T\frac{\mathbf{w}}{\|\mathbf{w}\|} \\ y(\mathbf{x}) &= 0 + r\|w\| \end{aligned} \tag{10}$$

or

$$r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}. \tag{11}$$

Notation: As with linear regression models, it is convenient to use a more compact notation in which we introduce an additional dummy 'input' value $x_0 = 1$ and define $\tilde{\mathbf{w}} = (w_0, \mathbf{w})$ and $\tilde{\mathbf{x}} = (x_0, \mathbf{x})$ so that

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{w}^T\mathbf{x} + w_0 \\ &= (w_0\ \mathbf{w}^T)\begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix} \\ &= \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}. \end{aligned} \tag{12}$$

4.1.2 Multiple classes

Now consider the extension of linear discriminants to $K > 2$ classes. We are tempted to build a K-class discriminator by combining a number of 2-class discriminators but this is not without problems.

One-versus-rest classifier: Consider the use of $K - 1$ classifiers each of which solves a two-class problem of separating points in class $\mathcal{C}_k$ from those not in $\mathcal{C}_k$ as in Fig. 7(a). This approach leads to regions of input space that are ambiguously classified.

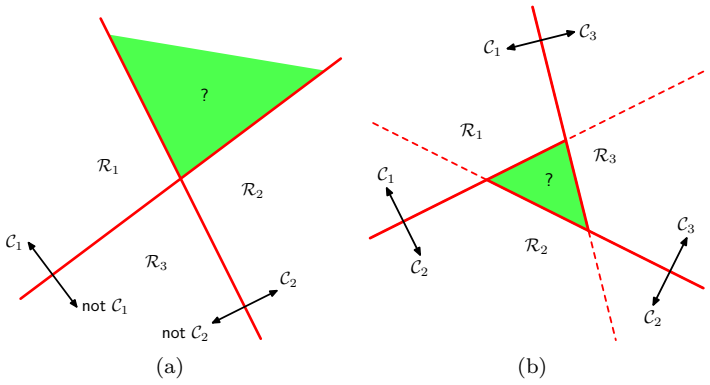

Figure 7:

One-versus-one classifier: An alternative is to introduce $K(K-1)/2$ binary discriminant functions, one for every possible pair of classes. Each point is then classified according to majority vote amongst the discriminant functions. This too leads to ambiguous regions as in Fig. 7(b).

Single K-class discriminator: We can avoid the above difficulties by considering a single K-class discriminant comprising K linear functions of the form

$$y_k(\mathbf{x}) \quad = \quad \mathbf{w}_k^T \mathbf{x} + w_{k0} \tag{13}$$

and then assigning a point $\mathbf{x}$ according to

$$\mathcal{C}_k \quad = \quad \arg \max_{1 \leq j \leq K} y_j(\mathbf{x}) \tag{14}$$

i.e. assign to the class for which the output y_k is largest. The decision boundary between class $\mathcal{C}_k$ and class $\mathcal{C}_j$ is therefore given by $y_k(\mathbf{x}) = y_j(\mathbf{x})$ and hence corresponds to a $(D-1)$-dimensional hyperplane defined by

$$(\mathbf{w}_k - \mathbf{w}_j)^T \mathbf{x} + (w_{k0} - w_{j0}) \quad = \quad 0. \tag{15}$$

The decision regions of such a discriminant are always singly connected and convex. We see in Fig. 8, that for any two points $\mathbf{x}_A$ and $\mathbf{x}_B$ in $\mathcal{R}_k$ any point along the line connecting these two points will also be in $\mathcal{R}_k$.

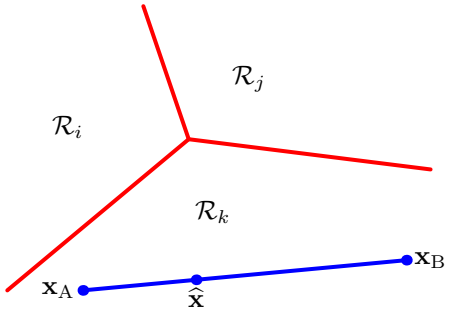

Figure 8:

What remains to be found in these linear models for classification are the model parameters, $\mathbf{w}_k$ which will establish the decision surfaces (hyperplanes). We will examine three approaches to finding $\mathbf{w}_k$: least squares, Fisher's linear discriminant, and the perceptron algorithm.