CS181 Section 3
Linear Classification

1 Linear Classifiers

The goal in classification is to take an input vector x and assign it to one of K discrete classes C_k where $k = 1, ..., K$. The input space is thus divided into **decision regions** whose boundaries are called **decision boundaries or surfaces**.

1.1 Discriminant Functions with Binary Responses

A discriminant function is one that directly assigns each vector x to a specific class. We first assume two classes, i.e. our responses are binary and $K = 2$. Linear classification seeks to divide the 2 classes by a linear separator in the feature space - if $d = 2$ the separator is a line; if $d = 3$ the separator is a plane; for general d the separator is a $(d-1)$-dimensional hyperplane.

The simplest representation of a linear discriminant function is obtained by taking a linear function of the input vector as such:

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

The corresponding decision boundary is defined by the relation $y(\mathbf{x}) = 0$, which corresponds to the $(d-1)$-dimensional hyperplane within the d-dimensional input space. $\mathbf{w}$ is orthogonal to every vector lying within the decision surface (prove this!) so $\mathbf{w}$ determines the orientation of the decision boundary. Furthermore, w_0 (called the **bias** or negative threshold), determines the location of the decision boundary.

$$\mathbf{w}^T(\mathbf{x}_A - \mathbf{x}_B) = 0$$

where $\mathbf{x}_A, \mathbf{x}_B$ both lie on the decision boundary.

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

In more compact notation, if $\tilde{\mathbf{w}} = (w_0, \mathbf{w})$ and $\tilde{\mathbf{x}} = (x_0, \mathbf{x})$, then

$$y(\mathbf{x}) = \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}$$

1.2 Fisher's Linear Discriminant

A useful way of thinking about linear classification is in terms of dimensionality reduction. We take a d-dimensional vector x and project it down into 1 dimension with $\mathbf{w}^T\mathbf{x}$. In the two-class problem where there are N_1 data points in C_1 and N_2 in C_2, we want to maximize the separation of the mean class vectors (to make it easier to find a linear separator):

$$m_2 - m_1 = w^T(\mathbf{m}_2 - \mathbf{m}_1)$$

as well as minimize total within-class variance for the whole dataset:

$$s_1^2 + s_2^2, \text{ where } s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$$

All this is encapsulated in the Fisher's criterion:

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$

which if we try to maximize results in the Fisher's linear discriminant (derivation in Bishop):

$$\mathbf{w} \propto \mathbf{S}_w^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

1.3 Perceptron Algorithm

Another well-known example of a linear discriminant model is the Perceptron Algorithm. It corresponds to a 2-class model where the input vector $\mathbf{x}$ is first transformed using a fixed non-linear transformation to give a feature vector $\phi(\mathbf{x})$ which is used to construct a generalized linear model of the form:

$$y(\mathbf{x}) = f(\mathbf{w}^T \phi(\mathbf{x}))$$

where $f()$ is an activation function (inverse of what is called link function in most statistics literature). The perceptron algorithm proposes an alternative error function known as the **perceptron criterion** given by:

$$E_P(\mathbf{w}) = -\sum_{n \in M} \mathbf{w}^T \phi_n t_n$$

M represents all the misclassified patterns. We then apply stochastic gradient descent to this error function, where the change in weight vector is given by:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_p(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta \phi_n t_n$$

η is the learning rate parameter and τ is an integer that indexes the steps of the algorithm. Note that as the weight vector evolves during training, the set of patterns that are misclassified will also change.

2 Practice Problems

1. **Hyperplanes and Discriminant functions**

   Suppose we have the discriminant function $y(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + w_0$, and if $y(\boldsymbol{x}) \geq 0$ we assign $\boldsymbol{x}$ to $\mathcal{C}_1$, and if $y(\boldsymbol{x}) < 0$ we assign $\boldsymbol{x}$ to $\mathcal{C}_2$. Show that for any x_0, x_1 on the decision boundary $(x_0 - x_1)$ is perpendicular to the vector $\boldsymbol{w}$.

2. **Convex Hulls and Linear Seperability**

   Define the convex hull of a set of data points ($\{\boldsymbol{x}_i\}$) as the set

   $$\left\{ \sum_i \alpha_i \boldsymbol{x}_i \mid \alpha_i \geq 0, \sum_i \alpha_i = 1 \right\}$$

   Additionally, define that two sets of points are linearly separable if there exists a vector $\boldsymbol{w}$ and w_0 such that $\boldsymbol{w}^\top \boldsymbol{x}_n + w_0 > 0$ for all points in the first set and $\boldsymbol{w}^\top \boldsymbol{y}_n + w_0 < 0$. Show that if two sets of points $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{y}_i\}$ are linearly separable, their convex hulls do not intersect.

3. **Perceptron Algorithm**

Consider the perceptron algorithm which is a binary classification algorithm that finds the best linear hyperplane to separate the basis-transformed input values. The error function that is minimized is 0 when the algorithm correctly labels a data point and otherwise:

$$E_p(\boldsymbol{w}) = -\sum_{n \in M} \boldsymbol{w}^\top \phi(\boldsymbol{x}_n) t_n,$$

where we sum over the mislabeled values and $t_n = 1$ if the correct classification is $\mathcal{C}_1$ and $t_n = -1$ if the correct classification is $\mathcal{C}_2$. Derive the Stochastic Gradient Descent relation to optimize the weight vector for this error function.

4. **Thresholded Discriminant Functions**

Suppose we have the discriminant function $y(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + w_0$, but that rather than assigning $\boldsymbol{x}$ to $\mathcal{C}_1$ when $y(\boldsymbol{x}) \geq 0$ and to $\mathcal{C}_2$ otherwise (as in Bishop 4.1.1), we instead assign $\boldsymbol{x}$ to $\mathcal{C}_1$ when $y(\boldsymbol{x}) \geq \eta$ for some η and to $\mathcal{C}_2$ otherwise. Do we gain any generality by moving to this thresholded decision rule? Why or why not?

5. **Maximizing Separation Between Classes (Bishop 4.4)**

Suppose, as in Fisher's Discriminant Analysis, that we want to find the vector $\boldsymbol{w}$ that maximizes the distance between the means of two classes $\mathcal{C}_1, \mathcal{C}_2$ that are projected onto it. That is, we want to maximize

$$\boldsymbol{w}^\top(\boldsymbol{m}_2 - \boldsymbol{m}_1), \tag{Bishop 4.2.2}$$

where $\boldsymbol{m}_k = \frac{1}{N_k}\sum_{\boldsymbol{x}\in\mathcal{C}_k} \boldsymbol{x}$.

(a) Show that by maximizing the criterion above subject to the constraint that $\boldsymbol{w}^\top\boldsymbol{w} = 1$, we find that $\boldsymbol{w}_{\max} \propto (\boldsymbol{m}_2 - \boldsymbol{m}_1)$. That is, $\boldsymbol{w}_{\max} = \alpha(\boldsymbol{m}_2 - \boldsymbol{m}_1)$ for some α.

(b) Geometrically, what is the interpretation of $\boldsymbol{w}_{\max}$?

6. **Fisher Criterion in Matrix Form (Bishop 4.5)**

The Fisher Criterion is defined as

$$J(\boldsymbol{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \tag{Bishop 4.2.5}$$

where

$$\begin{aligned} m_k &= \boldsymbol{w}^\top \boldsymbol{m}_k \\ \boldsymbol{m}_k &= \frac{1}{N_k}\sum_{\boldsymbol{x}\in\mathcal{C}_k} \boldsymbol{x} \\ s_k^2 &= \sum_{\boldsymbol{x}\in\mathcal{C}_k} (\boldsymbol{w}^\top\boldsymbol{x} - m_k)^2 \end{aligned}$$

Show that we can write $J(\boldsymbol{w})$ in matrix form as

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}},$$

where

$$\boldsymbol{S}_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top$$

and

$$\boldsymbol{S}_W = \sum_{\boldsymbol{x}\in\mathcal{C}_1} (\boldsymbol{x} - \boldsymbol{m}_1)(\boldsymbol{x} - \boldsymbol{m}_1)^\top + \sum_{\boldsymbol{x}\in\mathcal{C}_2} (\boldsymbol{x} - \boldsymbol{m}_2)(\boldsymbol{x} - \boldsymbol{m}_2)^\top$$

7. **Parsimonious models**

Softmax used in logistic regression is expressed as:

$$\Pr(t_k = 1 \mid \boldsymbol{x}, \{\boldsymbol{w}_{k'}\}_{k'=1}^{K}) = \frac{\exp\{\boldsymbol{w}_k^{\mathsf{T}}\boldsymbol{x}\}}{\sum_{k'=1}^{K}\exp\{\boldsymbol{w}_{k'}^{\mathsf{T}}\boldsymbol{x}\}}\,.$$

Show that the model for softmax is not parsimonious. That is, the solution w_k is not unique. Then, show how to add a contraint to make the model parsimonious.

8. **Classification with Same-Mean Different-Variance Gaussians (McKay 39.4)**

Consider the task of recognizing which of two Gaussian distributions a data point $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ comes from. We will assume that the two distributions have exactly the same mean but different variances. Let the probability that $\mathbf{x}$ is in class C_i (where $i \in \{0, 1\}$) be given by

$$\Pr(\mathbf{x}|C_i) = \prod_{j=1}^{D}\mathcal{N}(x_j|\mu_j, \sigma_{ij})$$

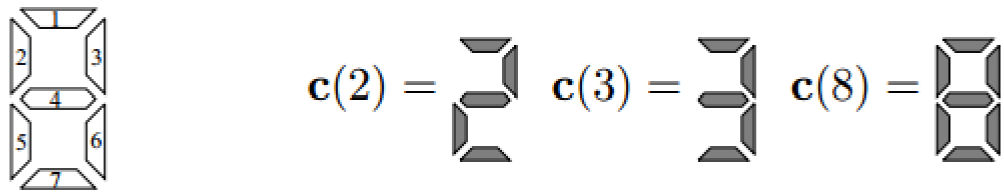
Show that $P(C_0|x)$ can be written in the form

$$P(C_0|x) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{y} + \theta)}$$

where y_i is an appropriate function of x_i, $y_i = g(x_i)$, and θ is some constant.

9. **LED with Errors (McKay 39.5)**

Consider an LED display with 7 elements numbered as shown below.



The state of the display is a vector $\mathbf{x}$. When the controller wants the display to show character number s, e.g. $s = 2$, each element x_j ($j \in \{1, 2, \ldots, 7\}$) either adopts its intended state $c_j(s)$ with probability $1 - f$ or is flipped with probability f. We will say $x_i = 1$ if the element is on and $x_i = 0$ if it isn't.

Assuming that 1) the LED displays an 8 (so that $x_i = 1$ for all $i \in \{1, 2, \ldots, 7\}$) and 2) you know that the true s was either a 2, 3, or 8 with prior probabilities p_2, p_3, p_8 respectively, what is the probability of $s = 8$. More specifically, compute $P(s = 8 | x_1 = 1, x_2 = 1, \ldots, x_7 = 1, s \in \{2, 3, 8\})$.

10. **Logistic sigmoid function (Bishop, 4.7)**

Show that the logistic sigmoid function

$$\sigma(a) = \frac{1}{1 + e^{-a}} \tag{1}$$

satisfies the property $\sigma(-a) = 1 - \sigma(a)$, and that it's inverse is given by

$$\sigma^{-1}(p) = \log \frac{p}{1 - p} \tag{2}$$

If we use $\sigma(a)$ to model a probability, p, what is an interpretation of the logistic inverse function?

11. **Exponential Family**

A distribution is part of the exponential family if we can rewrite its density function in the following way, given a parameter p and θ, a function of p:

$$f(x|p) = h(x)e^{\theta T(x) - A(\theta)}$$

Here, h and T are functions of x, and A is a function of θ. θ is known as the natural parameter of the distribution. Show that the Bernoulli distribution is part of the exponential family, and find its natural parameter. Recall the Bernoulli PMF:

$$f(x|p) = p^x(1-p)^{1-x}$$

How does the natural parameter relate to the logistic function given below, which we use for logistic regression to solve binary classification?

$$f(x) = \frac{e^x}{1+e^x}$$

12. **Margin distances**

Consider the hyperplane given by $\boldsymbol{w}^T\boldsymbol{x} + b = 0$. For an arbitrary data point $\boldsymbol{x}$, what is the distance between $\boldsymbol{x}$ and the hyperplane, in terms of $\boldsymbol{w}$ and b?