Statistical Machine Learning 2018

Exercises, week 8

9 November 2018

TUTORIAL

Exercise 1

Linear discriminant functions (Bishop, §4.1). Consider the discriminant function $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$, (where $\mathbf{w} \neq 0$). The decision surface is given by $y(\mathbf{x}) = 0$ (see figure below).

1. Consider the points $\hat{\mathbf{x}}$ on the decision surface, so $y(\hat{\mathbf{x}}) = 0$. We want to find the point $\hat{\mathbf{x}}^*$ that is closest to the origin. To find this point, minimize $||\hat{\mathbf{x}}||^2$ under the constraint $y(\hat{\mathbf{x}}) = 0$ using Lagrange multipliers, and show that the minimizing point $\hat{\mathbf{x}}^*$ satisfies

$$\hat{\mathbf{x}}^* = -\frac{w_0}{||\mathbf{w}||^2}\mathbf{w} \tag{1}$$

   So, the distance of the decision surface to the origin is $||\hat{\mathbf{x}}^*||$. Show that this distance is

$$||\hat{\mathbf{x}}^*|| = \frac{|w_0|}{||\mathbf{w}||} \tag{2}$$

Now consider an arbitrary point $\mathbf{x}$ and let $\mathbf{x}_\perp$ be its orthogonal projection onto the decision surface (implying $y(\mathbf{x}_\perp) = 0$), so that

$$\mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{||\mathbf{w}||} \tag{3}$$

2. Show that

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

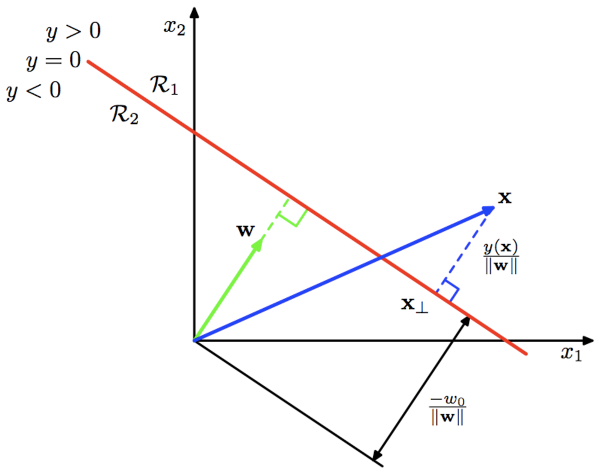

Figure 4.1 - Linear discriminant function in 2d.

Exercise 2

Fisher's linear discriminant (Bishop, §4.1.4). Consider two classes. Take an $\mathbf{x}$ and project it down to one dimension using

$$y = \mathbf{w}^T\mathbf{x}$$

Let the two classes have two means:

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{n\in C_1}\mathbf{x}^n \qquad \mathbf{m}_2 = \frac{1}{N_2}\sum_{n\in C_2}\mathbf{x}^n$$

We can choose $\mathbf{w}$ to maximize $\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)$, subject to $\sum_i w_i^2 = c$ where $c > 0$ is a constant. Show, using a Lagrange multiplier for the constraint (see appendix E), that this maximization leads to $\mathbf{w} \propto \mathbf{m}_1 - \mathbf{m}_2$

Exercise 3

Consider a binary classification problem. The two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ have a Gaussian class-conditional density, with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ resp. and shared covariance matrix $\boldsymbol{\Sigma}$. The prior class probabilities are $p(\mathcal{C}_1) = \pi$ and $p(\mathcal{C}_2) = (1 - \pi)$.

1. Is this a generative or discriminative probabilistic model? Why?

2. Show the posterior probability for class $\mathcal{C}_1$ can be written as linear discriminant function

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0) \tag{4}$$

with $\sigma(a)$ the logistic sigmoid, defined as

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \tag{5}$$

Hint: use $p(\mathcal{C}_1|\mathbf{x}) = \dfrac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$.

Suppose we have a data set $\{\mathbf{x}_n, t_n\}$ of N observations $\mathbf{x}$ with corresponding class labels t, where $t = 1$ denotes class $\mathcal{C}_1$ and $t = 0$ denotes class $\mathcal{C}_2$. We are looking for a maximum likelihood expression for the parameters in our model. Intuitively it is 'obvious' that, for example, the ML-solution for $\boldsymbol{\mu}_1$ should be given by the mean of all input vectors $\mathbf{x}_n$ of class $\mathcal{C}_1$

$$\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{\mathbf{x}_n \in \mathcal{C}_1} \mathbf{x}_n \tag{6}$$

with N_1 the number of data points belonging to class $\mathcal{C}_1$.

3. Show this intuition is valid by obtaining an expression for the likelihood of the dataset and then maximizing this w.r.t. $\boldsymbol{\mu}_1$.

BONUS PRACTICE

Exercise 4

Consider two basic patterns, represent by vectors $\mathbf{x}_0$ and $\mathbf{x}_1$. One of the two patterns (sequence of numbers) is transmitted over a noisy channel and received at the other end as pattern $\mathbf{y}$.
So, if the pattern that is being transmitted is $\mathbf{x}_s$ (with $s \in \{0, 1\}$), then the pattern that is received at the other end is a noisy version:

$$\mathbf{y} = \mathbf{x}_s + \mathbf{n}$$

where $\mathbf{n}$ is noise. The problem is to guess which pattern was transmitted: the pattern with $s = 0$ or the one with $s = 1$.

In a Gaussian channel, the noise is assumed to be distributed according to a zero-mean multivariate Gaussian,

$$p(\mathbf{n}|\mathbf{\Lambda}) = \mathcal{N}(\mathbf{n}|0, \mathbf{\Lambda}^{-1}) = \left|\frac{\mathbf{\Lambda}}{2\pi}\right|^{1/2} \exp\left(-\frac{1}{2}\mathbf{n}^T\mathbf{\Lambda}\mathbf{n}\right) \tag{7}$$

1. Show that the likelihood of receiving vector $\mathbf{y}$ given source $s \in \{0, 1\}$ is given by

$$p(\mathbf{y}|s) = \left|\frac{\mathbf{\Lambda}}{2\pi}\right|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{x}_s)^T\mathbf{\Lambda}(\mathbf{y} - \mathbf{x}_s)\right) \tag{8}$$

2. The optimal detector is based on the posterior probability ratio. Show that this ratio can be written as

$$\frac{p(s = 1|\mathbf{y})}{p(s = 0|\mathbf{y})} = \exp\left(\mathbf{y}^T\mathbf{\Lambda}(\mathbf{x}_1 - \mathbf{x}_0) + c\right) \tag{9}$$

   where c is a constant independent of the received pattern $\mathbf{y}$. Can you interpret each of the terms in the final expression?

3. Show this corresponds to a linear discriminant function $a(\mathbf{y}) = \mathbf{w}^T\mathbf{y} + w_0$ with decision boundary $a(\mathbf{y}) = 0$.

Exercise 5

Linear separation. (Exercise 4.1 in Bishop)

Given a set of data points $\{\mathbf{x}_n\}$, we can define the *convex hull* to be the set of all points $\mathbf{x}$ given by:

$$\mathbf{x} = \sum_n \alpha_n \mathbf{x}_n,$$

where $\alpha_n \geq 0$ and $\sum_n \alpha_n = 1$. Consider a second set of points $\{\mathbf{y}\}_n$ together with their corresponding convex hull. By definition, the two sets of points will be *linearly separable* if there exists a vector $\mathbf{w}$ and a scalar w_0 such that $\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0 > 0$ for all $\mathbf{x}_n$ and $\mathbf{w}^{\mathrm{T}}\mathbf{y}_n + w_0 < 0$ for all $\mathbf{y}_n$. Show that if their convex hulls intersect, the two sets of points cannot be linearly separable, and conversely that if they are linearly separable, their convex hulls do not intersect.