Statistical Machine Learning 2018

Exercises and answers, 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

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

   ANSWER: Lagrangian is

$$L = ||\hat{\mathbf{x}}||^2 + \lambda(\mathbf{w}^T\hat{\mathbf{x}} + w_0) = \sum_j \hat{x}_j^2 + \lambda(\sum_j w_j\hat{x}_j + w_0)$$

   Take partial derivative, and set zero: $\partial L/\partial\hat{x}_k = 0$ leads to

$$\hat{x}_k(c) = -\frac{\lambda}{2}w_k = cw_k$$

   where $c = -\lambda/2$. To find c^*, we plug $\hat{x}_k(c)$ into the constraints

$$0 = \sum_j w_j\hat{x}_k(c) + w_0 = c^*\sum_j w_j^2 + w_0 = c^*||\mathbf{w}||^2 + w_0$$

   so $c^* = -w_0/||\mathbf{w}||^2$ and $\hat{\mathbf{x}}^* = \hat{\mathbf{x}}(c^*)$, resulting in (1).
   Second part. Note that $||\alpha\mathbf{x}|| = |\alpha|\,||\mathbf{x}||$. So, fill in:

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

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

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

ANSWER:

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{w}^T\mathbf{x} + w_0 \\ &= \mathbf{w}^T(\mathbf{x}_\perp + r\frac{\mathbf{w}}{||\mathbf{w}||}) + w_0 \\ &= \mathbf{w}^T\mathbf{x}_\perp + w_0 + r\mathbf{w}^T\frac{\mathbf{w}}{||\mathbf{w}||} \\ &= y(\mathbf{x}_\perp) + r\frac{||\mathbf{w}||^2}{||\mathbf{w}||} \\ &= 0 + r||\mathbf{w}|| \end{aligned}$$

Divide left and right-hand side by $||\mathbf{w}||$, and the result follows.

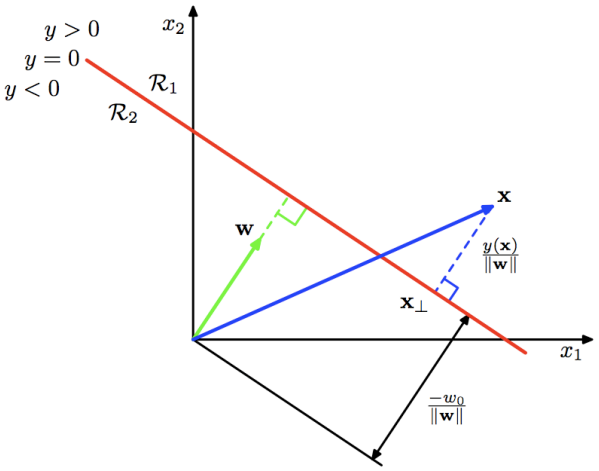

Figure 4.1 - Linear discriminant function in 2d.

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 $\mathbf{w} \propto \mathbf{m}_1 - \mathbf{m}_2$

ANSWER: Lagrangian is

$$L = \mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2) + \lambda(||\mathbf{w}||^2 - c) = \sum_j w_j(\mathbf{m}_1 - \mathbf{m}_2)_j + \lambda(\sum_i w_i^2 - c)$$

Derivative to w_j yields

$$(\mathbf{m}_1 - \mathbf{m}_2)_j + 2\lambda w_j = 0$$

So we have the result

$$w_j = \alpha(\mathbf{m}_1 - \mathbf{m}_2)_j$$

where $\alpha = -\frac{1}{2\lambda}$ a constant (which does need to be determined in this exercise).

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?

   ANSWER:
   It is a generative model as it tries to model the density of data $\mathbf{x}$ within each different class, which, together with prior class probabilities, could be used to generate data according to this distribution.

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

   ANSWER:
   Rewriting the hint in the form of the sigmoid (5) we have $p(\mathcal{C}_1|\mathbf{x}) = \frac{1}{1+\exp(-a)} = \sigma(a)$ with

   $$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \tag{6}$$

   Substituting the parameters for the Gaussian class-conditional densities into (6) this gives

   $$a = \ln \left\{ \frac{\exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)\right\}\pi}{\exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)\right\}(1 - \pi)} \right\}$$

   taking the log and writing out and collecting terms in the exponentials gives

   $$a = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{\pi}{(1 - \pi)}$$

   where the quadratic terms in $\mathbf{x}$ have cancelled. This is in the required form (4) with the first term linear in $\mathbf{x}$ and the last three independent of $\mathbf{x}$.

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 $\mathcal{C}_1$

$$\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{\mathbf{x}_n \in \mathcal{C}_1} \mathbf{x}_n \tag{7}$$

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. $\boldsymbol{\mu}_1$.

   ANSWER:
   The probability for a point $\{\mathbf{x}_n, \mathcal{C}_1\}$, with label $t_n = 1$, is given by $p(\mathbf{x}_n, \mathcal{C}_1) = \pi \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma})$. Similarly, $p(\mathbf{x}_n, \mathcal{C}_2) = (1 - \pi)\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Therefore the likelihood of the entire data set is given by

$$p(\mathbf{X}, \mathbf{t} | \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1 - \pi)\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1 - t_n} \tag{8}$$

   In order to maximize this w.r.t. $\boldsymbol{\mu}_1$ it is more convenient to maximize the log of the likelihood, and pick out only those terms that actually depend on $\boldsymbol{\mu}_1$, giving

$$\sum_{n=1}^{N} t_n \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = -\frac{1}{2} \sum_{n=1}^{N} t_n (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) + const$$

   Taking the derivative w.r.t. $\boldsymbol{\mu}_1$ (using C.19) and setting equal to zero gives

$$\begin{aligned} \sum_{n=1}^{N} t_n \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) &= 0 \\ \sum_{n=1}^{N} t_n \mathbf{x}_n &= \sum_{n=1}^{N} t_n \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_1 &= \frac{1}{N_1} \sum_{n=1}^{N} t_n \mathbf{x}_n \end{aligned}$$

   which corresponds to the mean of all input vectors belonging to class $\mathcal{C}_1\}$.

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,

$$p(\mathbf{n}|\mathbf{\Lambda}) = \mathcal{N}(\mathbf{n}|0, \mathbf{\Lambda}^{-1}) = \left|\frac{\mathbf{\Lambda}}{2\pi}\right|^{1/2} \exp\left(-\frac{1}{2}\mathbf{n}^T\mathbf{\Lambda}\mathbf{n}\right) \tag{9}$$

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

$$p(\mathbf{y}|s) = \left|\frac{\mathbf{\Lambda}}{2\pi}\right|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{x}_s)^T\mathbf{\Lambda}(\mathbf{y} - \mathbf{x}_s)\right) \tag{10}$$

   ANSWER: All you need to do is recognize that the difference $\mathbf{y} - \mathbf{x}_s$ corresponds (by definition) to the noise vector $\mathbf{n}$. Plugging this into (9) gives the desired result.

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

$$\frac{p(s = 1|\mathbf{y})}{p(s = 0|\mathbf{y})} = \exp\left(\mathbf{y}^T\mathbf{\Lambda}(\mathbf{x}_1 - \mathbf{x}_0) + c\right) \tag{11}$$

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

   ANSWER: Start by using Bayes and rewrite the posterior ratio as

$$\frac{p(s = 1|\mathbf{y})}{p(s = 0|\mathbf{y})} = \frac{p(\mathbf{y}|s = 1)p(s = 1)}{p(\mathbf{y}|s = 0)p(s = 0)} \tag{12}$$

   Plugging in the likelihood from (10) and collecting all terms in a single exponential gives

$$= \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{x}_1)^T\mathbf{\Lambda}(\mathbf{y} - \mathbf{x}_1) + \frac{1}{2}(\mathbf{y} - \mathbf{x}_0)^T\mathbf{\Lambda}(\mathbf{y} - \mathbf{x}_0) + \ln\frac{p(s = 1)}{p(s = 0)}\right) \tag{13}$$

   writing out and collecting terms then results in

$$= \exp\left(\mathbf{y}^T\mathbf{\Lambda}(\mathbf{x}_1 - \mathbf{x}_0) + c\right) \tag{14}$$

   with $c = -\frac{1}{2}\mathbf{x}_1^T\mathbf{\Lambda}\mathbf{x}_1 + \frac{1}{2}\mathbf{x}_0^T\mathbf{\Lambda}\mathbf{x}_0 + \ln\frac{p(s=1)}{p(s=0)}$, independent of $\mathbf{y}$.

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

   ANSWER: If we are trying to make an optimal decision ($s = 1$ or $s = 0$) for a given vector $\mathbf{y}$, then our best option is to guess the most probable pattern. Choosing

$$a(\mathbf{y}) = \mathbf{y}^T\mathbf{\Lambda}(\mathbf{x}_1 - \mathbf{x}_0) + c \tag{15}$$

then with $\mathbf{w} = \mathbf{\Lambda}(\mathbf{x}_1 - \mathbf{x}_0)$ and $w_0 = c$ this becomes the linear discriminant function $a(\mathbf{y}) = \mathbf{w}^T\mathbf{y} + w_0$, and

$$\begin{aligned} a(\mathbf{y}) > 0 &\Rightarrow s = 1 \\ a(\mathbf{y}) = 0 &\Rightarrow \text{boundary (guess either)} \\ a(\mathbf{y}) < 0 &\Rightarrow s = 0 \end{aligned}$$

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.

ANSWER: ($\Rightarrow$) If the two convex hulls intersect, then there exist α_n and β_n such that

$$\mathbf{x} = \sum_n \alpha_n \mathbf{x}_n = \sum_b \beta_n \mathbf{y}_n = \mathbf{y},$$

where $\alpha_n \geq 0, \beta_n \geq 0, \sum_n \alpha_n = 1, \sum_n \beta_n = 1$. If the two sets of points were linearly separable, then there would exist 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 $\mathbf{y}_n$. However, this would imply that $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \mathbf{w}^{\mathrm{T}}(\sum_n \alpha_n \mathbf{x}_n) + w_0 = (\sum_n \alpha_n \mathbf{w}^{\mathrm{T}}\mathbf{x}_n) + w_0 \overset{\sum_n \alpha_n = 1}{=} \sum_n \alpha_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + w_0) \overset{\alpha_n \geq 0}{>} 0$, while analogously $\mathbf{w}^{\mathrm{T}}\mathbf{y} + w_0 < 0$. Since $\mathbf{x} = \mathbf{y}$, the two statements are contradictory, so the two sets cannot be linearly separable.

($\Leftarrow$) If the two sets of points are linearly separable, then we have previously shown that $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 > 0$ and $\mathbf{w}^{\mathrm{T}}\mathbf{y} + w_0 < 0$, for some $\mathbf{w}, w_0$ and for any points $\mathbf{x}$ and $\mathbf{y}$ belonging to the convex hulls of $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_n\}$, respectively. This means that $\forall \mathbf{x}, \mathbf{y}$ we have $\mathbf{w}^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) > 0$, so $\mathbf{x} \neq \mathbf{y}$. Therefore, the two convex hulls do not intersect.