Statistical Machine Learning 2018

Exercises, week 9

16 November 2018

TUTORIAL

Exercise 1

We look at classification by the perceptron algorithm. The perceptron is an example of a linear discriminant model. It corresponds to a two-class model in which the input vector $\mathbf{x}$ is transformed into a feature vector $\boldsymbol{\phi}(\mathbf{x})$. This feature vector is then used to construct a linear model $y(\mathbf{x}) = f\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})\right)$, with bias component $\phi_0(\mathbf{x}) = 1$. The nonlinear activation function $f(\cdot)$ takes the form of a step function: $f(a) = +1$ for $a \geq 0$, $f(a) = -1$ for $a < 0$.

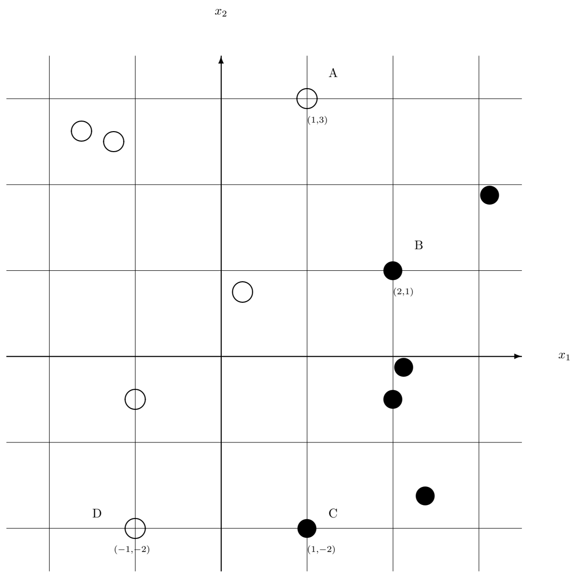

Figure 1: Data points from two classes

We use the $t \in \{-1, +1\}$ target coding scheme to let $t = +1$ denote class $\mathcal{C}_1$ and $t = -1$ denote class $\mathcal{C}_2$. Naturally, the problem now becomes how to determine the weight parameters $\mathbf{w}$ from a given dataset $\{\mathbf{x}_n, t_n\}$ in order to obtain the optimal classification scheme.

1. Explain how the perceptron classifies a (new) data point for a given set of $\mathbf{w}$. Why is it not a good idea to try to learn $\mathbf{w}$ simply from the error function defined by the total number of misclassified patterns in a data set (i.e. $E(\mathbf{w}) = \frac{1}{2}\sum_n |y(\mathbf{x}_n) - t_n|$) ?

A better approach is to define an error function known as the *perceptron criterion*

$$E_{\mathrm{P}}(\mathbf{w}) = -\sum_{n\in\mathcal{M}} \mathbf{w}^T\boldsymbol{\phi}_n t_n \tag{1}$$

with the sum taken over all misclassified patterns. We can use this function in a *stochastic gradient descent* technique: $\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} - \eta\nabla E_n$, with η a learning rate parameter.

2. Show this results in the following perceptron learning algorithm

$$\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} + \eta\boldsymbol{\phi}(\mathbf{x}_n)t_n \tag{2}$$

3. Show that in the perceptron learning algorithm, we can set the learning parameter η equal to 1 without loss of generality.

In Figure 1 a dataset of two classes is depicted: the solid circles belong to class $\mathcal{C}_1$ and the open circles belong to class $\mathcal{C}_2$. We will only look at the subset of points {A,B,C,D}. The features correspond directly to the parameters x_1 and x_2, i.e. $\boldsymbol{\phi}(\mathbf{x}) \equiv [1, x_1, x_2]$.

3. Assume $\mathbf{w}^{(0)} = [0, 1, 0]$. Which of the data points {A,B,C,D} is not classified correctly for this initial weight vector?

4. From the given $\mathbf{w}^{(0)}$, iterate over the set of points {A,B,C,D} (in that order) until the perceptron learning algorithm (2) reaches convergence (take $\eta = 1$). What is the final set of weight parameters?

Exercise 2

In *logistic regression* we start from the general form of the posterior probability of class $\mathcal{C}_1$, as a logistic sigmoid

$$\sigma(a) = \frac{1}{1+\exp(-a)} \tag{3}$$

acting on a linear function of the feature vector $\boldsymbol{\phi}$.

$$p(\mathcal{C}_1|\boldsymbol{\phi}) = y(\boldsymbol{\phi}) = \sigma(\mathbf{w}^T\boldsymbol{\phi}) \tag{4}$$

1. From the definition in (3), show that

$$\frac{d\ln\sigma}{da} = (1-\sigma) \tag{5}$$

2. For a data set $\{\boldsymbol{\phi}_n, t_n\}$, with $t_n \in \{0,1\}$, the likelihood function can be written as

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\{1-y_n\}^{1-t_n} \tag{6}$$

where $y_n = p(\mathcal{C}_1|\boldsymbol{\phi}_n)$. Define the *cross entropy* error as $E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w})$. Use the result (5) to show that the gradient of this error function w.r.t. $\mathbf{w}$ is given by

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\boldsymbol{\phi}_n \tag{7}$$

3. Describe how this can be used to obtain a gradient descent algorithm to learn the weight vector $\mathbf{w}$.

4. Show that if the classes are linearly separable, then the magnitude of $\mathbf{w}$ of the ML solution is unbounded (i.e. goes to infinity).

Exercise 3

Laplace approximation (Bishop §4.4). Consider the probability density $p(t)$ defined as

$$p(t) = \frac{1}{Z} f(t) \tag{8}$$

with

$$f(t) = \begin{cases} t^2 \exp(-\lambda t) & , \quad t \geq 0 \\ 0 & , \quad t < 0 \end{cases} \tag{9}$$

in which λ is a positive constant. Z is an (unknown) normalizing constant.

1. Show that the mode t^* of $p(t)$ is located at $t = 2/\lambda$.
2. Create a second order Taylor expansion of $\ln f(t)$ around $t^* > 0$. Take the exponential to obtain an approximation to $f(t)$ in the form of an unnormalized Gaussian with mode t^*.
3. Rewrite the approximation into a standard Gaussian $q(t)$, and use the normalization factor to compute an estimate for the unknown constant Z in (9). Compare to the true value of $Z = 2$ for $\lambda = 1$.

Exercise 4

Consider a data set $\mathcal{D}$ and a set of models $\{\mathcal{M}_i\}$ with parameters $\{\boldsymbol{\theta}_i\}$. In a Bayesian selection of the 'best' model, we would like to compare the posterior distribution of the models given the data: $p(\mathcal{M}_i|\mathcal{D})$. In doing so, the so called *model evidence*, quantifying the preference of the data for different models, plays an important role. From Bayes' theorem this is given by (eq.4.136)

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}_i) p(\boldsymbol{\theta}|\mathcal{M}_i) \mathrm{d}\boldsymbol{\theta} \tag{10}$$

1. Show why (and under what assumptions) the model evidence is a good measure for comparing the model posteriors $p(\mathcal{M}_i|\mathcal{D})$. (see Bishop, §3.4)

Using Laplace approximation (§4.4) we found a general result for the approximation of the normalizing constant of a distribution

$$Z = \int f(\mathbf{z}) \mathrm{d}\mathbf{z} \simeq f(\mathbf{z}_0) \int \exp\left\{ -\frac{1}{2} (\mathbf{z} - \mathbf{z}_0)^{\mathrm{T}} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) \right\} \mathrm{d}\mathbf{z} = f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \tag{11}$$

2. Match this with (10) to derive the following approximation to the log model evidence

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2} \ln 2\pi - \frac{1}{2} \ln |\mathbf{A}| \tag{12}$$

   Explain how the last three terms ('Occam factor', eq.4.137) penalize model complexity.