Reading Notes of Pattern Classification and Machine Learning

Tianyi Cui

August 12, 2012

1 Introduction

Different kinds of tasks of machine learning:

- supervised learning: known input and target vectors
- classification: output is one of a finite number of discrete categories
  - regression: output is one or more continuous variables
- unsupervised learning: no corresponding target values
  - clustering: discover groups of similar examples within the data
  - density estimation: determine the distribution of data within the input space
  - dimension reduction
- reinforcement learning: finding suitable actions to take in a given situation in order to maximize a reward

1.1 Example: Polynomial Curve Fitting

In regression problems, we can use a polynomial function

$$y(x, \mathbf{w}) = w_0 + w_1 + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1.1}$$

to fit the underlying function.

We need to minimize the *error function*

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1.2}$$

in which unique solution $\mathbf{w}^*$ can be found in closed form.

The root-mean-square (RMS) is error defined by

$$E_{RMS} = \sqrt{2E(\mathbf{w}^*)/\mathbf{N}} \tag{1.3}$$

When M is large, *over-fitting* occurs, i.e. E_{RMS} against test data becomes large. One technique to control over-fitting is *regularization*, by adding a penalty term to the error function (1.2) in order to discourage the coefficients from reaching large values:

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1.4}$$

1.2 Probability Theory

Equations for probability:

Sum rule

$$p(X) = \sum_Y p(X, Y) \tag{1.5}$$

Product rule

$$p(X, Y) = p(Y|X)p(X) \tag{1.6}$$

Bayes' theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{1.7}$$

The denominator in (1.7) can be expressed in terms of the quantities appearing in the numerator

$$p(X) = \sum_Y p(X|Y)p(Y) \tag{1.8}$$

We can view the denominator in Bayes' theorem as being the normalization constant required to ensure that the sum of the conditional probability on the left-hand side of (1.7) over all values of Y equals 1.

Before any observation, we have a probability of a certain event Y, this is called *prior probability* $p(Y)$, after some observation X, the probability of event Y becomes the *posterior probability* $p(Y|X)$.

X and Y are said to be *independent* if $p(X, Y) = p(X)p(Y)$, which is equivaent to $P(Y|X) = p(Y)$.

1.2.1 Probability densities

If the probability that x will lie in (a, b) is given by

$$p(x \in (a, b)) = \int_a^b p(x)\mathrm{d}x \tag{1.9}$$

then $p(x)$ is called the *probability density* over x.

Apparently $p(x) \geq 0$ and $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$.

Under a nonlinear change of variable, a probability density transforms differently from a simple function, due to the Jacobian factor. If $x = g(y)$, since $p_x(x)\mathrm{d}x = p_y(y)\mathrm{d}y$, hence

$$\begin{aligned} p_y(y) &= p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= p_x(g(y))\left|g'(y)\right| \end{aligned} \tag{1.10}$$

The *cuculative distribution function*

$$P(z) = \int_{-\infty}^{z} p(x)\mathrm{d}x \tag{1.11}$$

For several continuous variables $x_1, \ldots, x_D$, denoted collectively by the vector $\mathbf{x}$, then we can define a joint probability density $p(\mathbf{x})$ such that $p\left(\mathbf{x} \in (\mathbf{x}_0, \mathbf{x}_0 + \delta\mathbf{x})\right) = p(\mathbf{x}_0)\delta\mathbf{x}$.

1.2.2 Expectations and covariances

The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the *expectaion* of $f(x)$ and denoted by $\mathbb{E}[f]$. For a descrite distribution,

$$\mathbb{E}[f] = \sum_x p(x)f(x) \tag{1.12}$$

For continuous variables,

$$\mathbb{E}[f] = \int p(x)f(x)\mathrm{d}x \tag{1.13}$$

In either case, the expectation can be approximated given N samples,

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n) \tag{1.14}$$

When considering expectations of functions of several variables, we use subscript to indicate which variable is being averaged over, e.g. $\mathbb{E}_x[f(x,y)]$ is a function of y.

We can also consider *conditional expectation*

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x) \tag{1.15}$$

The *variance* of $f(x)$ is defined by

$$\begin{aligned}\mathrm{var}[f] &= \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)]\right]^2 & (1.16)\\ &= \mathbb{E}\left[f(x)^2\right] - \mathbb{E}\left[f(x)\right]^2 & (1.17)\end{aligned}$$

The *covariance* of two random variable x and y is defined by

$$\begin{aligned}\mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right]\\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] & (1.18)\end{aligned}$$

which expresses the extent to which x and y vary together. If they are independent, then thei covariance vanishes.

In the case of two vectors of random variables $\mathbf{x}$ and $\mathbf{y}$, the covariance is a matrix

$$\begin{aligned}\mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{x,y}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\left\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\right\}\right]\\ &= \mathbb{E}_{x,y}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}] & (1.19)\end{aligned}$$