Reading Notes of Pattern Classification and Machine Learning

Tianyi Cui

August 14, 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) error is 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:

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

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 *cumulative distribution function*

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

For several continuous variables $x_1, \dots, x_D$, 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

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

1.2.3 Bayesian probabilities

In the *classical* or *frequentist* interpretation of probability, probabilities is viewed in terms of the frequencies of random, repeatable events. In the more general *Beyesian* view, probabilities provide a quantification of uncertainty, so we can say the probability of an uncertain event, like whether the Arctic ice cap will have disappeared by the end of the century, which is not events that can be repeated.

In the polynomial curve fitting example, we assume the parameters $\mathbf{w}$ have a prior probability distribution $p(\mathbf{w})$, then given the observed data $\mathcal{D}$, the posterior probability is

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{1.20}$$

where the quantity $p(\mathcal{D}|\mathbf{w})$ is called the *likelihood function*, which expresses how probable the observed data set is for different settings of the parameter vector $\mathbf{w}$. The likelihood is not a probability distribution over $\mathbf{w}$, and its integral does not necessarily equal one.

Given the definition of liklihood, we can state Bayes' theorem in words

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{1.21}$$

where all of these quantities are viewed as functions of $\mathbf{w}$.

In the likelihood function $p(\mathcal{D}|\mathbf{w})$, in the frequentist setting, $\mathbf{w}$ is considered to be a fixed parameter, whose value is determines by some form of 'estimator', and error bars on this estimate are obtained by considering the distribution of possible data sets $\mathcal{D}$. By contrast, from Bayesian viewpoint there is only a single data set $\mathcal{D}$ (the one actually observed), and the uncertainty in the parameters is expressed through a probability distribution over $\mathbf{w}$.

1.2.4 The Gaussian distribution

The Gaussian distribution on a single real-valued variable x is defined by

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.22}$$

which is governed by two parameters: μ the *mean* and σ^2 the *variance*. σ is called the *standard deviation*, and $\beta = 1/\sigma^2$ is called the *precision*. The mean of x is given by $\mathbb{E}[x] = \mu$ and the variance of x is given by $\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$.

The Gaussian distribution defined over a D-dimensional vector $\mathbf{x}$ of continuous variables is given by

$$\mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1.23}$$

Suppose we have a data set of observation $\mathsf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$ which is *independent and identically distributed* (often abbreviated to i.i.d.) from a Gaussian distribution. The likelihood of the data set, which is a function of μ and σ^2, is in the form

$$p\left(\mathbf{x}|\mu,\sigma^2\right) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2) \tag{1.24}$$

One common criterion for determining the parameters in a probability distribution using an observed data set is to find the parameter values that maximize the likelihood function.

In practice, for mathematical and numerical reasons, it's more convenient to maximize the log of the likelihood functions

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{1.25}$$

Maximizing (1.25) with respect to μ gives the maximum likelihood solution

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1.26}$$

which is the *sample mean.* Similarly, Maximize (1.25) with respect to σ^2gives

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \tag{1.27}$$

which is the *sample variance.*

The maximum likelihood approach systematically underestimates the variance of the distribution. This is an example of a phenomenon called *bias* and is related to the problem of over-fitting encountered in the context of polynomial curve fitting. First, we note that μ_{ML} and σ^2_{ML} are functions of the data set values $x_1,\ldots,x_N$. Consider the expectations of these quantities with respect to the data set values, which themselves come from a Gaussian distribution with parameters μ and σ^2

$$\begin{aligned} \mathbb{E}[\mu_{\mathrm{ML}}] &= \mu & (1.28)\\ \mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \left(\frac{N-1}{N}\right)\sigma^2 & (1.29) \end{aligned}$$

so on average the maximum likelihood approach will underestimate the true variance by a factor $(N-1)/N$.

From (1.29) we see the following estimate for the variance parameter is unbiased

$$\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2 \tag{1.30}$$

this result arises automatically when we adopt a Bayesian approach (Section 10.1.3).

1.2.5 Curve fitting re-visited

The goal of curve fitting problem is to make predictions for the target variable t given some new value of the input variable x on the basis of a set of training data $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$ and $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$. We can express out uncertainty over the value of the target variable using a probability distribution. Assume that, given the value of x, the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, \mathbf{w})$ given by (1.1). Thus we have

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \tag{1.31}$$

where β is the precision parameter.

Use the training data $\{\mathbf{x}, \mathbf{t}\}$ to determine the values of the unknown parameters $\mathbf{w}$ and β by maximum likelihood, the likelihood function is

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right) \tag{1.32}$$

and its logarithm is

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \tag{1.33}$$

Maximizing (1.33) with respect to $\mathbf{w}$ gives us $\mathbf{w}_{\mathrm{ML}}$, which is the same as minimize the *sum-of-squares error function* defined by (1.2).

Maximizing (1.33) with respect to β gives

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathrm{w}_{\mathrm{ML}}) - t_n\}^2 \tag{1.34}$$

Having determined the parameters $\mathbf{w}$ and β, we can now make predictions for new values of x, and in probabilistic model, these are expressed in terms of the *predictive distribution* that gives the probability distribution over t

$$p(t|x, \mathrm{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{ML}), \beta_{\mathrm{ML}}^{-1}\right) \tag{1.35}$$

In a more Bayesian approach, we introduce a Gaussian prior distribution over the polynomial coefficients $\mathbf{w}$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\} \tag{1.36}$$

where α is the precision of the distribution and $M + 1$ is the number of elements in $\mathbf{w}$. Values such as α, which controls the distribution of model parameters, are called *hyperparameters.*

Using Bayes' theorem

$$p(\mathbf{w}|\mathsf{\mathbf{x}}, \mathsf{\mathbf{t}}\ , \alpha, \beta) \propto p(\mathsf{\mathbf{t}}\ |\mathsf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) \tag{1.37}$$

We can now determine $\mathbf{w}$ by finding the most probable value of $\mathbf{w}$ given the data, in other words by maximizing the posterior distribution. This technique is called *maximum posterior*, or simply *MAP*.

Taking the negative logarithm of (1.37) and combining with (1.33) and (1.36), we find that the maximum of the posterior is given by the minimum of

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \tag{1.38}$$

Thus we see that maximizing the posterior distribution is equivalent to minimizing the regularized sum-of-squares error function (1.4).

1.2.6 Bayesian curve fitting

Although we have included a prior distribution $p(\mathbf{w}|\alpha)$, we are still making a point estimate of $\mathbf{w}$ and so this does not yet amount to a Bayesian treatment. In a fully Bayesian approach, we should consistently apply the sum and product rules of probability, which requires, as we shall see shortly, that we integrate over all values of $\mathbf{w}$. Such marginalizations lie at the heart of Bayesian methods for pattern recognition.

In curve fitting, we are given the training data $\{\mathsf{\mathbf{x}}, \mathsf{\mathbf{t}}\}$, along with a new test point x, and our goal is to predict the value of t. Assuming the parameters α and β are fixed and known in advance by now, we wish the evaluate the predictive distribution $p(t|\mathsf{\mathbf{x}}, \mathsf{\mathbf{t}})$. Using the product rules of probability

$$p(t|x, \mathsf{\mathbf{x}}, \mathsf{\mathbf{t}}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathsf{\mathbf{x}}, \mathsf{\mathbf{t}})\mathrm{d}\mathbf{w} \tag{1.39}$$

Here $p(t|x, \mathbf{w})$ and $p(\mathbf{w}|\mathsf{\mathbf{x}}, \mathsf{\mathbf{t}})$ are given by (1.31) and normalizing the right-hand side of (1.37).

The calculation and integration in (1.39) can be performed analytically with the result in a Gaussian distribution

$$p(t|x, \mathsf{\mathbf{x}}, \mathsf{\mathbf{t}}) = \mathcal{N}\left(t|m(x), s^2(x)\right) \tag{1.40}$$

where the mean and variance are given by

$$\begin{aligned} m(x) &= \beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n && (1.41)\\ s^2(x) &= \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\phi}(x) && (1.42)\end{aligned}$$

Here the matrix $\mathbf{S}$ is given by

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x)^{\mathrm{T}} \tag{1.43}$$

and we have defined the vector $\boldsymbol{\phi}(x)$ with elements $\phi_i(x) = x^i$ for $i = 0, \ldots, M$.

The matrix and the mean of the predictive distribution in (1.40) is dependent on x. The first term in (1.42) represents the uncertainty due to the noise on the target variables, and the second term arises from the uncertainty in the parameters $\mathbf{w}$ and is a consequence of the Bayesian treatment.

1.3 Model Selection

Model selection is to find the appropriate values of complexity parameters within a given model and to find the best model for a particular application.

Due to the problem of over-fitting, performance on the training set is not a good indicator of predictive performance. If data is plentiful, we can set aside a *validation set* for comparing models. If the model design is iterated many times using a limited size data set, some over-fitting to the validation data can occur so it may be necessary to keep aside a third *test set* on which the performance of the selected model is finally evaluated.

But the supply of data for training and testing will be limited. To use as much of the available data as possible for training, one solution is to use *cross-validation*, which is, to divide the data into S sets, and use $S - 1$ sets for training and 1 set for validation, in total S runs. When $S = N$, it's called the *leave-one-out* technique.

One major drawback of cross-validation is that the number of training runs is increased by a factor of S, and this can be problematic when training is computationally expensive. And when there are multiple parameters to explore, required number of training runs is exponential in the number of parameters. We therefore need a measure of performence which depends only on the training data (i.e. not validation-based) and which does not suffer from bias due to over-fitting.

Historically various 'information criteria' have been proposed that attempt to correct for the bias of maximum likelihood by the addition of a penalty term to compensate for the over-fitting of more complex models. For example, the *Akaike information criterion*, or AIC, chooses the model for which the quantity

$$\ln p(\mathcal{D}|\mathbf{w}_{\mathrm{ML}}) - M \tag{1.44}$$

is largest. Later we'll see how complexity penalties arise in a natural and principled way in a fully Bayesian approach.

1.4 The Curse of Dimensionality

In the polynomial curve fitting example we had just one input variable x, but in practice we will deal with spaces of high dimensionality comprising many input variables. This poses some serious challenges and is an important factor influencing the design of pattern recognition techniques.

For example, a simple approach for classification is to divide the input space into regular cells and classify each cell independently. But the number of cells grows exponentially with the dimensionality of the space, so we need exponentially large quantity of training data in order to ensure that the cells are not empty, which is not practical in a space of more than a few variables. High-dimensional general polynomial curve fitting have similar problems, as D the number of input variables increases, the number of independent coefficients grows proportionally to D^M for a polynomial of order M.

Our geometrical intuitions formed from life can fail badly when we consider spaces of higher dimensionality. For example, consider a sphere of radius $r = 1$ in a space of D dimensions, the fraction of the volume of the sphere that lies between radius $r = 1 - \epsilon$ and $r = 1$ is given by $1 - (1 - \epsilon)^D$. For large D, this fraction tends to 1 even for small values of ϵ. Thus, in spaces of high dimensionality, most of the volume of a sphere is concentrated in a thin shell near the surface!

Similarly, consider Gaussian distribution in high-dimensional space. If we transform from Cartesian to polar coordinates, and then integrate out the directional variables, we obtain an expression for the density $p(r)$ as a function of radius r from the origin. We can see that for large D the probability mass of the Gaussian is concentrated in a thin shell.

The severe difficulty that can arise in spaces of many dimensions is sometimes called the *curse of dimensionality.* But it does not prevent us from finding effective techniques applicable to high-dimensional spaces. First, real data will often be confined to a region of the space having lower effective dimensionality, and in particular the directions over which important variations in the target variables occur may be so confined. Second, real data will typically exhibit some smoothness properties (at least locally) so that for the most part small changes in the input variables will produce small changes in the target variables, and so we can exploit local interpolation-like techniques to allow us to make predictions of the target variables for new values of the input variables. Successful pattern recognition techniques exploit one or both of these properties. For example, an application in manufacturing in which images are captured of identical planar objects on a conveyor belt, in which the goal is to determine their orientation. Each image is a point in a space whose dimensionality is determined by the number of pixels. But since there are three degrees of freedom of variability between images, actually a set of images will live on a three dimensional *manifold* embedded within the high-dimensional space.