Lecture Outline

Reading: Chapter 1

- Review of probability theory
- Bayesian probabilities
- Maximum Likelihood (ML) estimation

1.2 Probability Theory

1.2.1 Probability Densities

In addition to RVs which can only take on discrete values, we shall also consider RVs which are continuous. For a *probability density function* (pdf), $p(x)$, the probability that x will lie in an interval (a, b) is given by

$$p(x \in (a, b)) = \int_a^b p(x)dx. \tag{1}$$

We require a pdf to satisfy

$$p(x) \geq 0 \tag{2}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1. \tag{3}$$

The probability that x lies in the interval $(-\infty, z)$ is given by the *cumulative probability distribution function* (cdf)

$$P(z) = \int_{-\infty}^{z} p(x)dx. \tag{4}$$

The sum and product rules for continuous RVs take the form

$$p(x) = \int p(x, y)dy \tag{5}$$

$$p(x, y) = p(y|x)p(x) \tag{6}$$

If we have several continuous variables $(x_1, x_2, \ldots, x_D)$ denoted collectively by the vector $\mathbf{x}$, i.e. random vector or multidimensional RV, then we denote the joint pdf as $p(\mathbf{x}) = p(x_1, x_2, \ldots, x_D)$ with the same requirements as the scalar RV.

Under a nonlinear change of variable, a pdf transforms differently than a simple function. If x is a deterministic variable and x is transformed by $x = g(y)$, then $f(x)$ becomes $f(g(y))$, i.e. function composition. On the other hand if x and y are a RVs with pdfs $p_x(x)$ and $p_y(y)$ and x is transformed by $x = g(y)$ then $p_y(y) = p_x(g(y))|g'(y)|$ (note the Jacobian factor). For more information see Prob. 1.4 and online solution.

1.2.2 Expectations and Covariances

The average value of some function $f(x)$ under a pdf $p(x)$ is called the *expectation* of $f(x)$ and denoted $\mathbb{E}[f]$. For discrete RVs, we have

$$\mathbb{E}[f] = \sum_x p(x)f(x) \tag{7}$$

so that the average is weighed by the relative probabilities of the values of x. For continuous RVs we have

$$\mathbb{E}[f] = \int p(x)f(x)dx. \tag{8}$$

We can also consider *conditional expectation* with respect to a conditional distribution, so that

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x) \text{ discrete} \tag{9}$$

$$\mathbb{E}_x[f|y] = \int p(x|y)f(x)dx \text{ continuous.} \tag{10}$$

In either case, if we are given a finite number, N of points drawn by trial, then the expectation can be approximated as

$$\mathbb{E}[f] \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n). \tag{11}$$

We will use the notation $\mathbb{E}_x[f(x,y)]$ to indicate the average of the function $f(x,y)$ with respect to the distribution of x which results in a function only of y.

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

$$\begin{aligned}
\text{var}[f] &= \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])^2\right] \\
&= \mathbb{E}\left[f(x)^2 - 2f(x)\mathbb{E}\left[f(x)\right] + \mathbb{E}\left[f(x)\right]^2\right] \\
&= \mathbb{E}\left[f(x)^2\right] - 2\mathbb{E}\left[f(x)\right]\mathbb{E}\left[f(x)\right] + \mathbb{E}\left[f(x)\right]^2 \\
&= \mathbb{E}\left[f(x)^2\right] - \mathbb{E}\left[f(x)\right]^2
\end{aligned} \tag{12}$$

and provides a measure of the variability in $f(x)$ around its mean value $\mathbb{E}[f(x)]$. For the variance of x itself, i.e. $f(x) = x$, we have

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2. \tag{13}$$

For two RVs x and y, the *covariance* is defined by

$$\text{cov}[x,y] = \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \tag{14}$$

$$= \mathbb{E}_{x,y}\left[xy\right] - \mathbb{E}[x]\mathbb{E}[y] \tag{15}$$

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

In the case of vector RVs $\mathbf{x}$ and $\mathbf{y}$, the covariance matrix is given by

$$\text{cov}[\mathbf{x},\mathbf{y}] = \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\left\{\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T]\right\}\right] \tag{16}$$

$$= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\mathbf{x}\mathbf{y}^T\right] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T] \tag{17}$$

If we consider the covariance of the elements of the random vector $\mathbf{x}$ with each other, i.e. auto covariance, then we simplify the notation

$$\text{cov}[\mathbf{x}] = \text{cov}[\mathbf{x},\mathbf{x}]. \tag{18}$$

1.2.3 Bayesian Probabilities and Maximum Likelihood Estimators

The *classicial* or *frequentist* interpretation of probability views probabilities in terms of the frequencies of random, repeatable events.

The *Bayesian* interpretation of probability views probabilities as a quantification of uncertainty. Using the Bayesian interpretation, we would like to be able to quantify our expression of uncertainty and make precise revisions of uncertainty in the light of new evidence, as well as subsequently to be able to take optimal actions or decision as a consequence.

Reconsidering the polynomial curve fitting problem of finding $\mathbf{w}$ for data $\mathcal{D}$, we capture our assumptions about $\mathbf{w}$ before observing the data in the form of a prior probability distribution $p(\mathbf{w})$. The uncertainty in $\mathbf{w}$ *after* we have observed $\mathcal{D}$ is expressed in the form of the posterior probability $p(\mathbf{w}|\mathcal{D})$ using Bayes' theorem

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}. \tag{19}$$

Important Information: The term $p(\mathcal{D}|\mathbf{w})$ is evaluated for the observed data set $\mathcal{D}$ and can be viewed as a function of the parameter vector $\mathbf{w}$, in which case it is called the *likelihood function*. It expresses how probable the observed data set is for different settings of the parameter vector $\mathbf{w}$. Note that the likelihood is not a pdf over $\mathbf{w}$ and its integral with respect to $\mathbf{w}$ is not necessarily equal to one.

The denominator in (19) is a normalization constant, which ensures that the posterior distribution is a valid pdf. From (19), we also have (integrate both sides with respect to $\mathbf{w}$) from the sum and product rules

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w} \tag{20}$$

$$= \int p(\mathbf{w}, \mathcal{D})d\mathbf{w} \tag{21}$$

Definition: The *maximum likelihood* (ML) estimator chooses $\mathbf{w}$ such that $p(\mathcal{D}|\mathbf{w})$ is maximized. The *error function* is the negative log of the likelihood function. Maximizing the likelihood is equivalent to minimizing the error, i.e.

$$\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w}) \Leftrightarrow \min_{\mathbf{w}} \left[-\log p(\mathcal{D}|\mathbf{w})\right] \tag{22}$$

since $-\log$ is monotonically decreasing.

Notes:

- Equation (22) helps us optimally find $\mathbf{w}$ for data $\mathcal{D}$
- Usually $\mathbf{w}$ is a statistical model
- We can solve (22) in closed form if $p(\mathbf{w})$ has a normal or uniform distribution