Multivariate Probability and Statistics

Jussi Martin

October 24, 2016

Introduction

This presentation is a brief summary of some of the topics introduced in chapter 4 of the book *Natural Image Statistics* (reference at the last slide).

Since there won't be enough time to cover the whole chapter 4, I decide to go trough the things that are needed for understanding the *Bayes' rule*, which I think is one of the core contents in this chapter.

I will also introduce *expectation*, *variance*, *covariance*, *likelihood*, *log-likelihood*, *maximum likelihood estimate* and *maximum a posteriori estimates* in these slides, since some of them are needed in the exercises.

Random Variables

Discrete random variable is a variable which can have any value from some given countable set with some given probabilty. Example: outcome of throwing a dice. The outcome can have any integer value from 1 to 6, each with the probability $\frac{1}{6}$. Continuos random variable is a variable which can have any value from some given continuous range of values, with some given probabilities for the value been in any given interval. Example: uniform probability distribution on the interval $[0, 1]$. Now the outcome is in an interval $[a, b]$ with the probability $b - a$ for any given values $0 \leq a \leq b \leq 1$.

Random Vectors

Random vector is a n-tuple of random variables. Each of these can be either similar or different types of variables. Formally:

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$$

where $\mathbf{z}$ is the random vector and the random variables z_i are its components.

Probability Density Function (pdf)

Let $P(z \text{ is in [a, b]})$ be the probability that the value of a random variable z is in the interval $[a, b]$, then the probabilty density function $p_z(a)$ can be defined approximately as follows:

$$p_z(a) \approx \frac{P(z \text{ is in } [a, a + \nu])}{\nu}$$

for very small values of ν.
Similarily, if **z** is a random vector:

$$p_{\mathbf{z}}(\mathbf{a}) \approx \frac{P(z_i \text{ is in } [a_i, a_i + \nu] \text{ for all } i)}{\nu^n}$$

for very small values of ν. Rigorous definitions are obtained by taking the one-sided limit $\nu \to 0^+$.

Joint and Marginal pdfs

The pdf $p(z_1, z_2, \ldots, z_n)$ of a random vector $\mathbf{z}$ is also called joint pdf, since it depends on all the components z_i.
The marginal pdfs $p_{z_i}(z_i)$ are obtained by integrating over the other variables (that is z_j, with $j \neq i$). Example in two dimensions:

$$p_{z_1}(z_1) = \int p_{\mathbf{z}}(z_1, z_2) dz_2.$$

Conditional Probabilities

Conditional pdfs are obtained from any joint pdf by fixing the value of one (or several) of its components and multiplying the result with a normalization factor. Example: (conditional pdf of z_2 with z_1 given)

$$p(z_2|z_1 = a) = \frac{p_{\mathbf{z}}(a, z_2)}{\int p_{\mathbf{z}}(a, z_2) dz_2}.$$

Using the definition of marginal pdf, this can be written as

$$p(z_2|z_1 = a) = \frac{p_{\mathbf{z}}(a, z_2)}{p_{z_1}(a)},$$

which can be simplified further by omiting the subcsripts

$$p(z_2|z_1 = a) = \frac{p(a, z_2)}{p(a)}$$

Conditional Probabilities

or by not introducing the quantity a we can write it as

$$p(z_2|z_1) = \frac{p_{\mathbf{z}}(z_1, z_2)}{p(z_1)}.$$

For the discrete case the integral is replaced by sum, that is

$$P(z_2|z_1) = \frac{P_{\mathbf{z}}(z_1, z_2)}{P_{z_1}(z_1)}$$

where

$$P_{z_1}(z_1) = \sum_{z_2} P_{\mathbf{z}}(z_1, z_2).$$

Independence

Two random variables z_1 and z_2 are said to be statistically independent if information about the value of one of them does not give any information about the value of the other. Formally this can be written as

$$p(z_2|z_1) = p(z_2) \text{ for every } z_1 \text{ and } z_2,$$

since conditional probability assumes that the value of the other variable is fixed (and hence known). Substituting the formula of $p(z_1|z_2)$ we obtain two alternative ways to define independence:

$$\frac{p(z_1, z_2)}{p(z_1)} = p(z_2) \text{ for every } z_1 \text{ and } z_2$$

or

$$p(z_1, z_2) = p(z_1)p(z_2) \text{ for every } z_1 \text{ and } z_2.$$

Bayes' Rule

Bayes' rule gives us the probability distribution $p(\mathbf{s}|\mathbf{z})$ when we know the distributions $p(\mathbf{z}|\mathbf{s})$ and $p(s)$. For continuously distributed random vector it is written as follows:

$$p(\mathbf{s}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{s})p_{\mathbf{s}}(\mathbf{s})}{\int p(\mathbf{z}|\mathbf{s})p_{\mathbf{s}}(\mathbf{s})ds}.$$

The distribution on the left is called *posterior distribution* and the distribution $p_{\mathbf{s}}(\mathbf{s})$ on the right side is called *prior distribution*.
In the case of discretely distributed random vector the rule takes the form:

$$P(\mathbf{s}|\mathbf{z}) = \frac{P(\mathbf{z}|\mathbf{s})P_{\mathbf{s}}(\mathbf{s})}{\sum_{\mathbf{s}} P(\mathbf{z}|\mathbf{s})P_{\mathbf{s}}(\mathbf{s})}.$$

Expectation

Expectation $E\{z\}$ of a random variable z is a weighted average of the outcomes which the variable can obtain, with the weights been the individual propabilities (or values of the pdf in the continuous case).

$$\text{Discrete case:} \qquad E\{z\} = \sum_z P_z(z)z$$

$$\text{Continuous case:} \qquad E\{z\} = \int p_z(z)zdz$$

The expectation can be also defined for random vectors componentwise:

$$E\{\mathbf{z}\} = \begin{pmatrix} E\{z_1\} \\ E\{z_2\} \\ \vdots \\ E\{z_n\} \end{pmatrix}$$

where $E\{z_i\}$ is defined as above, with z replaced by z_i for all i.

Variance and Covariance

Variance in some sense tells how close the probability mass is concentrated near the expected value. It is defined as follows:

$$var(z) = E\{z^2\} - (E\{z\})^2.$$

Alternatively it can be written as

$$var(z) = E\{(z - E\{z\})^2\}.$$

When we have two variables we can define their covariance:

$$cov(z_1, z_2) = E\{z_1 z_2\} - E\{z_1\}E\{z_2\}$$

which measures how well we can predict the value of one variable from another by using simple linear predictor.
Covariance is often normalized and this quantity is the correlation coefficient:

$$corr(z_1, z_2) = \frac{cov(z_1, z_2)}{\sqrt{var(z_1)var(z_2)}}.$$

Likelihood and Log-likelihood

Given n samples $z(1), \ldots, z(n)$ of a random variable variable z, which is assumed to have probability distribution $p(z, \alpha)$ for some unknown parameter α, we can define the likelihood:

$$p(z(1), \ldots, z(n)|\alpha) = p(z(1)|\alpha) \times \ldots \times p(z(n)|\alpha)$$

and the log-likelihood:

$$\log p(z(1), \ldots, z(n)|\alpha) = \log p(z(1)|\alpha) + \ldots + \log p(z(n)|\alpha).$$

These are functions of the parameter α, when the model $p(z, \alpha)$ and the samples $z(1), \ldots, z(n)$ are fixed.

Maximum Likelihood and Maximum a Posteriori

In maximum likelihood and maximum a posteriori estimates one tries to find the value of the parameter α which is in some sense optimal for observing the samples $z(1), \ldots, z(n)$.
Using the Bayes' rule we see that

$$p(\alpha|z(1), \ldots, z(n)) = \frac{p(z(1), \ldots, z(n)|\alpha)p(\alpha)}{p(z(1), \ldots, z(n))}.$$

The maximum a posteriori estimate $\hat{\alpha}$ is the one that maximizes the value of the posterior probability for the given prior $p(\alpha)$.
Maximum likelihood estimate $\hat{\alpha}$ on the other hand is the one which maximizes the likelihood.
If we choose a flat prior, that is $p(\alpha) =$ constant, these estimates are the same.

References

Aapo Hyvärinen, Jarmo Hurri, and Patrik O. Hoyer
Natural Image Statistics: A Probabilistic Approach to Early Computational Vision
Springer-Verlag, 2009