Lecture 4: Conditional probability and correlation

Recap

Lecture 3 examined transformation from one probability density function to another and also joint probabiility density functions. We ended by paving the way for looking a conditional probability. This lecture will examine conditional probability density function in more detail and then look at correlation.

We finished up by writing out formal definitions for conditional probability:

$$F_x(r\,|s) = \text{probability that } r < x \le r + dr \text{ given that } y = s. \tag{1}$$

When we count points in a given bin, we can say that out of N points total, the bin defined by $r < x \le r + dr, s < y \le s + ds$ will contain $NF_{xy}(r, s)\, dr\, ds$ points. If we consider a slice defined by $s < y \le s + ds$, for any value of r, it will contain $NF_y(s)\, ds$ points. The fraction in $r < x \le r\, dr$ given that $y = s$ is

$$F_x(r\,|s)\; dr = \frac{NF_{xy}(r,s)\; dr\; ds}{NF_y(s)\; ds} \tag{2}$$

and the conditional pdf is

$$F_x(r\,|s\,) = \frac{F_{xy}(r,s)}{F_y(s)} \tag{3}$$

Bayes' Theorem

The formal definition for conditional probability can be written for r in terms of s, or for s in terms of y. We have

$$F_x(r\,|s\,) = \frac{F_{xy}(r,s)}{F_y(s)} \tag{4}$$

and also

$$F_y(s\,|r\,) = \frac{F_{xy}(r,s)}{F_x(r)} \tag{5}$$

We can combine these in a number of ways:

$$F_x(r\,|s\,) = \frac{F_y(s|r)F_x(r)}{F_y(s)} \tag{6}$$

Equivalently:

$$F_x(r\,|s\,) = \frac{F_y(s|r)F_x(r)}{\int_{-\infty}^{\infty} F_{xy}(r,s)\,dr} = \frac{F_y(s|r)F_x(r)}{\int_{-\infty}^{\infty} F_y(s|r)F_x(r)\,dr} \tag{7}$$

This expression is called Bayes' Theorem and provides a formal framework for considering the probability of an event given prior knowledge.

If the random variables x and y are independent, then $F_x(r|s)$ is independent of s, which implies from (4) that

$$F_{xy}(r,s) = F_x(r)F_y(s). \tag{8}$$

General form of joint Gaussian pdf

To place the formal definitions in context, consider a joint pdf for independent Gaussian variables:

$$F_{xy}(r,s) = \frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(\frac{-r^2}{2\sigma_x^2}\right)\frac{1}{\sqrt{2\pi}\sigma_y}\exp\left(\frac{-s^2}{2\sigma_y^2}\right) \tag{9}$$

$$= \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[\frac{-1}{2}\left(\frac{r^2}{\sigma_x^2}+\frac{s^2}{\sigma_y^2}\right)\right]. \tag{10}$$

If x and y are uncorrelated, the joint pdf is either isotropic (if $\sigma_x = \sigma_y$) or has no tilt.

We can write the joint Gaussian distribution in a general form for a collection of variations $x_1, x_2, ...x_N$, with $\sigma_1 = \sigma_2 = 1$:

$$F_{x_1x_2....}(r_1, r_2...) = (2\pi)^{-N/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{N} r_i^2\right] \tag{11}$$

Of course things change if we have two correlated variables, and in class we looked at the joint pdf that emerges from correlated noise, for example when x is drawn from a Gaussian distribution and $y = x + r$, where r is noise drawn from a Gaussian distribution. We also looked at the correlation of y and z when $z = x + s$, where s is different from r and also drawn from a Gaussian distribution. Both cases result in a tilted joint pdf, providing clear evidence that x and y (or x and z) are correlated.

```
%  define correlated noise
x=randn(100000,1); y=randn(100000,1)+x;
z=randn(100000,1)+x;

% plot joint pdf for x and y
histogram2(x,y,'Normalization','pdf','Displaystyle','tile')

% plot joint pdf for y and z
histogram2(y,z,'Normalization','pdf','Displaystyle','tile')
```

Covariance

Calculating the joint pdf is often more than we can accomplish from real data. The **covariance** is a simple statistic relating variables x and y:

$$C_{xy} = \langle x'y' \rangle, \tag{12}$$

where the primes indicate that these are fluctuations about the mean. The covariance of a variable with itself is the **variance**:

$$C_{yy} = \langle y'y' \rangle. \tag{13}$$

The **correlation** is sort of a normalized covariance:

$$\rho_{xy} = \frac{\langle x'y' \rangle}{\sqrt{\langle x'^2 \rangle \langle y'^2 \rangle}}. \tag{14}$$

How can we interpret the correlation. Let's consider a linear model, where y is a linear function of x. In the following, we assume that variables x and y zero means, or equivalently that they have had their means removed, so the primes are dropped. A linear relationship between modeled $\hat{y}$ and measured x is

$$\hat{y}' = \alpha x', \tag{15}$$

where α is a constant chosen to make $haty$ approximate y.

We could also write this in a more general form as a matrix equation to fit lots of coefficients α_j to multiple form of data. In general form, we would write

$$y_i = \sum_{j=1}^{N} A_{ij}\alpha_j, \tag{16}$$

where the elements of A_{ij} represent the jth element of data type i. As a matrix equation we would write

$$\mathbf{y} = \mathbf{A}\alpha, \tag{17}$$

where $\mathbf{y}$ is a vector with M elements, α is a vector with N elements, and $\mathbf{A}$ is an $M \times N$ matrix. We'll come back to this case later.

Let's continue with the one variable fit that we're considering now. We choose to minimize the mean-square error (mse):

$$\epsilon = \langle(\hat{y} - y)^2\rangle = \alpha^2\langle x^2\rangle - 2\alpha\langle xy\rangle + \langle y^2\rangle. \tag{18}$$

The best α in the sense that the mse is minimized is found by differentiating with respect to α, setting the result equal to zero, and solving for α. Because $\epsilon \to \infty$ as $\alpha \to \pm\infty$, the result is a minimum.

$$\frac{\partial\epsilon}{\partial\alpha} = 2\alpha\langle x^2\rangle - 2\langle xy\rangle = 0. \tag{19}$$

Thus:

$$\alpha = \frac{\langle xy\rangle}{\langle x^2\rangle} \tag{20}$$

The term α is a regression coefficient, and it assumes a fully linear relationship between x and y.

If we plug α into the equation for the mse, we can find the misfit

$$\begin{aligned}
\epsilon &= \alpha^2\langle x^2\rangle - 2\alpha\langle xy\rangle + \langle y^2\rangle &(21)\\
&= \frac{\langle xy\rangle^2}{\langle x^2\rangle} - 2\frac{\langle xy\rangle^2}{\langle x^2\rangle} + \langle y^2\rangle &(22)\\
&= \langle y^2\rangle\left(1 - \frac{\langle xy\rangle^2}{\langle x^2\rangle\langle y^2\rangle}\right) &(23)\\
&= \langle y^2\rangle\left(1 - \rho_{xy}^2\right) &(24)
\end{aligned}$$

Thus the mean-squared error (the mse) is related to the variance of the quantity that we were trying to fit ($\langle y^2\rangle$) multipled by 1 minus the correlation coefficient squared.