

Probability Review

Vassil Chatalbashev

1/9/03

1 Probability Distributions

$P_X(X = x)$, often abbreviated as just $P(x)$.

Facts:

$$0 \leq P(x)$$
$$\sum_x P(x) = 1$$

1.1 Marginal Distributions

$$P_X(X = x)$$

$$P_{Y,X}(X = x, Y = y)$$

$$P_{Y,X,Z}(X = x, Y = y, Z = z)$$

Marginalization Rule (summing out):

$$P_X(X = x) = \sum_{y \in Y} P_{X,Y}(X = x, Y = y)$$

1.2 Conditional Distributions

Incorporate some prior information, e.g. $X = x$

Rule:

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

Recall from above that $P(Y) = \sum_x P(X = x, Y)$

1.3 Bayes Rule

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}$$

Can be derived from the above rule for conditional distributions.

1.4 Independence

X is independent of Y ($I(X, Y)$) if:

$$P(X|Y) = P(X)$$

i.e. Y gives us no additional information.

Equivalently:

$$I(X, Y) \Leftrightarrow P(X, Y) = P(X)P(Y)$$

2 Probability Mass Functions vs. Probability Density Functions

For discrete random variables, we have a *Probability Mass Function (PMF)*, whereas for continuous, we have a *Probability Density Function (PDF)*.

Essentially *probability mass* is actual probability, whereas *probability density* needs to be integrated over some region to yield a probability mass.

When we have a continuous variable, it has infinitely many possible values, so the probability that it takes on a particular value is 0. That's why for continuous r.v.'s we look at the probability of intervals.

Let $f(x)$ be the PDF, for a continuous r.v. X , then:

$$P(a < X < b) = \int_a^b f(x)dx$$

As expected, we must have:

$$f(x) \geq 0$$
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

All of the above probability rules work for both discrete and continuous r.v.'s, except that marginalization will now use integration instead of summation.

3 Expectation, Variance

The expectation of an r.v. X is:

$$E(x) = \sum_{x \in X} P(x)x$$

or

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx$$

for continuous variables.

Similarly we can have an expectation of a function of a random variable. It is easy to see that expectation is a linear operator, i.e. $E(g(x) + h(x)) = E(g(x)) + E(h(x))$.

The variance of an r.v. X is:

$$\sigma^2 = \sum_{x \in X} (x - E(X))^2 P(x)$$

Standard deviation σ is the square root of the variance.

Covariance of two r.v.'s X and Y is:

$$\sigma^2 = \sum_{x \in X, y \in Y} (x - E(X))(y - E(Y))P(x, y)$$

4 Some important distributions

4.1 Discrete:

Bernoulli:

$$X \in 0, 1$$
$$P(X = 1) = p$$
$$P(X = 0) = 1 - p$$

The only parameter of the distribution is p . Note that instead we can write:

$$P(X = x) = p^x(1 - p)^{1-x}$$

Check what happens for $x \in 0, 1$.

4.2 Continuous:

Gaussian(Normal):

$$X \in \mathfrak{R}$$
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where the parameters are $\mu = E(X)$, $\sigma^2 = Var(X)$

Multivariate Gaussian:

$$X \in \mathfrak{R}^n$$
$$f(x) = \frac{1}{2\pi^{n/2}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Here $\mu \in \mathfrak{R}^n$, so $\mu_i = E(X_i)$. $\Sigma \in \mathfrak{R}^{n \times n}$ is the covariance matrix. $\Sigma_{i,j} = E(X_i X_j) - \mu_i \mu_j$. So diagonal entries are variance, and off-diagonal entries are the co-variance between the different entries x_i of the vector X . $|\Sigma|$ denotes the determinant of Σ , and Σ^{-1} is the inverse.

5 Fitting Maximum Likelihood Parameters

Suppose we have a set of examples $D = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$. We assume they are generated from a certain distribution (say Bernoulli), and wish to find the parameters of the distribution.

We write the probability of all examples, assuming they are i.i.d (independent and identically distributed.):

$$P(D) = \prod_1^m P(x_i)$$
$$L(D) = \sum_1^m \log P(x_i)$$

In writing the probability of D the i.i.d assumption enabled us to simply multiply the individual probabilities (independence) and also represent them using the same distribution $P(x_i)$ (identically distributed.)

We then took the logarithm of the total probability in order to make differentiation later easy. All we now need to do is to plug in the actual distribution we are assuming in place of $P(x_i)$. Then we can differentiate the log-likelihood w.r.t. to the distribution parameters (for example for Bernoulli, we would just differentiate w.r.t to $p = P(X = 1)$), to find the parameters that maximize it.