

Statistics Review

Machine Learning 2005

**Todd K. Leen
Machine Learning**

Statistics Review

- **In order to sharpen up our language, we need to agree on some notions from probability and statistics**

Some Probability and Statistics

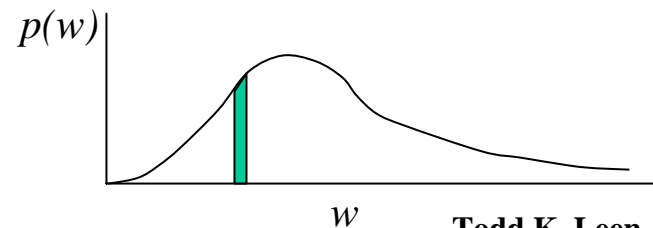
- Discrete (continuous) random variables are characterized by their probability distribution (density function).
 - Discrete Y takes on one of m values. $Pr(Y=y_i)=Pr(y_i)$ is the probability that Y takes on the value y_i on any given trial.

e.g. Flip a coin, what's probability of a heads?
 - Continuous X in R^N ,

$p(x) d^N x$ is probability that X takes on a value in the volume element $d^N x$ centered at the value x .

e.g. What's the probability that an adult human weighs between 110 and 112 lbs?

$$P(110 \leq W \leq 112) = \int_{110}^{112} p(w) dw$$



Todd K. Leen
Machine Learning

Some Probability and Statistics

- **Normalization** -- the probability that you take a measurement and get some value is unity

- Discrete
$$\sum_{i=1}^m P(y_i) = 1$$

- Continuous
$$\int_{-\infty}^{\infty} p(x) dx = 1$$

- **Expected values.** Let f be a function of a random variable. The expected value of f is

- Discrete
$$E[f] = \sum_{i=1}^m P(y_i) f(y_i)$$

- Continuous
$$E[f] = \int_{-\infty}^{\infty} p(x) f(x) dx$$

Some Probability and Statistics

- **Expected values** (cont'd). If f is the identity, then

- Discrete $E[Y] \equiv \langle Y \rangle \equiv \bar{Y} = \sum_{i=1}^m P(y_i) y_i \equiv \mu_Y$

- Continuous $E[x] = \int_{-\infty}^{\infty} p(x) x dx$

- **Variance** $var(x) = E[(x-E[x])^2] = E[x^2] - (E[x])^2$

Multivariate Random Variables

$$x \in \mathbb{R}^n, \quad p(x_1, \dots, x_n) \geq 0$$

$$\begin{aligned} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p(x_1, \dots, x_n) dx_1 \dots dx_n \\ \equiv \int p(x) d^n x = 1 \end{aligned}$$

Marginal density e.g. on x_1 $p_1(x_1) = \int p(x_1, \dots, x_n) dx_2 \dots dx_n$

A set of random variables are **statistically independent** if their joint density is the product of the marginals

$$p(x_1, x_2, \dots, x_n) = p_1(x_1) p_2(x_2) \dots p_n(x_n)$$

Multivariate Random Variables

Conditional density: The probability density for x_1 , **given** the value for x_2 is called the **conditional** probability and written

$$P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)}$$

If x_1 and x_2 are **independent**, then

$$P(x_1|x_2) = \frac{P_1(x_1)P_2(x_2)}{P_2(x_2)} = P_1(x_1)$$

e.g. multivariate Gaussian

$$P(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

where the **mean** vector is

$$\mu \equiv E[x]$$

and the **covariance** matrix is

$$\Sigma \equiv E[(x-\mu)(x-\mu)^T]$$

Gaussian Random Variables (cont'd)

$$p(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

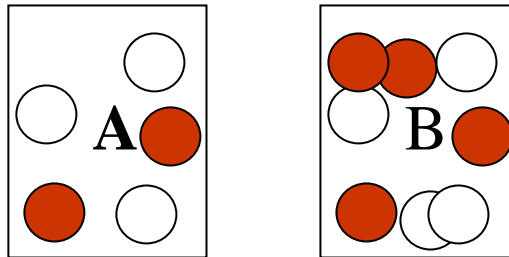
Note if the covariance is **diagonal** (and since it's a real, symmetric matrix, there's **always** a coordinate system in which the covariance matrix is diagonal) then the density breaks into a product of the marginals:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \sigma_3^2 & \\ 0 & & & \ddots \\ & & & & \sigma_n^2 \end{pmatrix} \longrightarrow p(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right]$$
$$\equiv p_1(x_1)p_2(x_2)\dots p_n(x_n)$$

Thus, Gaussian random variables that are uncorrelated (diagonal covariance) are also statistically independent -- though this is **not true** for more general distributions.

Joint and Conditional Probabilities

- Illustrative example – Two boxes (A and B) each contain red and white balls. Box A has r_A red balls and w_A white balls. Similarly for box B .



- I reach into box A and pull out a ball. The probability that it's red is

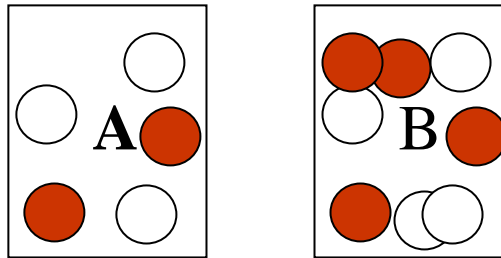
$$p(r | A) = r_A / (r_A + w_A) \quad (\text{prob. of } \underline{\text{red given}} \text{ box } A)$$

similarly for box B $p(r | B) = r_B / (r_B + w_B)$

Conditional and Joint Probabilities

- Joint density

I pick one of the boxes at random $p(A)+p(B)=1$, and draw a ball from that box.



What's the probability that I picked box *A* and I drew a red ball

$$p(r,A) = p(r | A) p(A) \text{ is the joint density}$$

so the conditional probability is $p(r | A) = p(r,A)/p(A)$

Note normalization $p(r | A) + p(w | A) = 1$.

Marginal Probability

- Let's define a *color* variable $c \in \{r, w\}$ and a *box* variable $b \in \{A, B\}$
- I select one of the boxes at random, and pick a ball from it at random. The probability distribution on the color is

$$p(c) = p(c | A) p(A) + p(c | B) p(B) = \sum_b p(b) p(c | b) = \sum_b p(c, b)$$

the last equality says we get the marginal probability distribution by summing over the other variables (one in this case) in the joint probability distribution.

- This holds for continuous random variables with the sum replaced by an integral.

Bayes' Theorem

- Bayes' theorem is a probabilistic inversion.
Take the expressions for the joint probability

$$p(c, b) = p(c | b) p(b)$$

and

$$p(c, b) = p(b | c) p(c)$$

and conclude

$$p(b|c) = \frac{p(c | b) p(b)}{p(c)} = \frac{p(c | b) p(b)}{\sum_b p(c, b)} = \frac{p(c | b) p(b)}{\sum_b p(c | b) p(b)}$$

Bayes' Theorem

- We started by quantifying the probability of choosing a ball of color c given a particular box $p(c|b)$.

Bayes' theorem tells us how to invert this. Given a particular color ball, what's the probability that it came from box b ?

$$p(b|c) = \frac{p(c|b) p(b)}{p(c)}$$

So I hand you a ball that came from one of the boxes (I don't tell you which box). You observe the color, and you can tell me the probability that it came from box A .

Estimation

- Usually we don't have the densities, but rather a bunch of data -- and we will often build **estimates** (e.g. of expectations) based on the data.

e.g. estimating the mean from a string of scalar values, the data \mathbf{D}

$$\mathbf{D} = (x^1, x^2, \dots, x^N), \quad x^I \in \mathbb{R}$$

The usual estimator of the mean is the sample mean $\hat{\mu} = \frac{1}{N} \sum_{I=1}^N x^I$

This is a random variable, because if I re-do the experiment of selecting N samples of x from the population (defined by the density $p(x)$), I'll get a different number for the sample mean. What's its expectation?

$$E[\hat{\mu}] = \int \left(\frac{1}{N} \sum_{I=1}^N x^I \right) p(x^1, \dots, x^N) d^N x$$

Estimation

We have
$$E[\hat{\mu}] = \int \left(\frac{1}{N} \sum_{i=1}^N x^i \right) p(x^1, \dots, x^N) d^N x$$

To evaluate this, we usually make two assumptions about the sample density

1. Each x_i is independent of the others $p(x_1, x_2, \dots) = p_1(x_1) p_2(x_2) \dots$
2. Each carries the same probability density $p_i(x_i) = p(x_i)$

(The samples are said to be independent and identically distributed, or iid for short.)

Thus
$$p(x^1, x^2, \dots, x^N) = p(x^1) p(x^2) \dots p(x^N)$$

Given this assumption, the expectation above becomes

$$E[\hat{\mu}] = \frac{1}{N} \sum_{i=1}^N E[x^i] = E[x] = \mu$$

The expected value of the sample mean is the population mean. The sample mean is thus an **unbiased estimator** of the population mean.

Estimation

- We've seen that the sample mean has expected value equal to the population mean. Its variance is given by (show this)

$$\begin{aligned}\text{Var } \hat{\mu} &= E[\hat{\mu}^2] - E[\hat{\mu}]^2 \\ &= E\left[\left(\frac{1}{N} \sum_{i=1}^N x_i^2\right)^2\right] - \mu^2 \\ &= \frac{1}{N^2} (N(\sigma_x^2 + \mu^2) + N(N-1)\mu^2) - \mu^2 \\ &= \frac{1}{N} \sigma_x^2\end{aligned}$$