Basic Mathematics for Statistical Inference

Zengchang Qin

Intelligent Computing and Machine Learning Lab
School of ASEE, Beihang University
zengchang.qin@gmail.com

September 27, 2011

Abstract

This is a summary of some basic mathematics that are commonly used in machine learning and data mining.

1 Combinational Analysis

Combinatorics is a branch of mathematics concerning the study of finite or countable discrete structures. $\binom{n}{r}$ represents the number o possible combinations of n objects take r at a time.

$$\binom{n}{r} = \frac{n!}{(n-r)!r!} \tag{1}$$

The binomial theorem use

$$(x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r} \tag{2}$$

2 Probability

The probability a discrete variable x takes value X is: $0 \leq P(x = X) \leq 1$. Let $\mathcal{X}$ be the set of all possible values of x, we then have:

$$\sum_{X' \in \mathcal{X}} P(x = X') = 1$$

If x is a continuous variable[1]:

$$\int p(x)dx = 1$$

$P(x = X, y = Y)$ is the **joint probability** that both event $x = X$ and $y = Y$ occur. Variables can be "summed out" of joint distributions, this is called **marginal distribution**, i.e.,

$$p(x) = \int_y p(x,y)dy$$

or

$$p(y) = \int_x p(x,y)dx$$

If $P(x,y) = P(x)P(y)$, then random variable x and y are **independent**. If and only if x and y are independent, the equation $P(x,y) = P(x)P(y)$ holds. Otherwise:

$$P(x,y) = P(x|y)P(y)$$

[1] We use $P(\cdot)$ to represent the discrete probability and $p(\cdot)$ to represent distribution.

where $P(x|y)$ is called **conditional probability**. Similarly, the following equation also holds

$$P(x,y) = P(y|x)P(x)$$

We then can obtain:

$$P(x|y)P(y) = P(y|x)P(x)$$

By moving $P(y)$ to the right-hand side of the equation, we will then get one of the most important equation in modern probability theory: Bayes rule or **Bayes theorem**:

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

It can be also written in another form:

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{y\in\mathcal{Y}} P(y|x)P(x)}$$

where $\mathcal{Y}$ is the set of all possible values of y. Sometimes, it is also written as:

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{y'} P(y'|x)P(x)}$$

Here, we are free to condition the whole thing on any set of assumptions, $\mathcal{H}$, such that:

$$\sum_y P(x,y|\mathcal{H}) = P(x|\mathcal{H})$$

The Bayes rule then becomes:

$$P(x|y,\mathcal{H}) = \frac{P(y|x,\mathcal{H})P(x|\mathcal{H})}{P(y|\mathcal{H})}$$

Bayesian methods for machine learning is becoming popular in recent years. We can rewrite Bayes theorem in the machine learning style: Given training data $\mathcal{D}$ and a model $\mathcal{H}$ (e.g., linear model, neural networks or other probabilistic models which are characterized by parameters) defined by parameter θ:

$$P(\theta|\mathcal{D},\mathcal{H}) = \frac{P(\mathcal{D}|\theta,\mathcal{H})P(\theta|\mathcal{H})}{P(\mathcal{D}|\mathcal{H})}$$

where $P(\mathcal{D}|\theta,\mathcal{H})$ is the **likelihood**, $P(\theta|\mathcal{H})$ is called **prior probability** and $P(\theta|\mathcal{D},\mathcal{H})$ is called **posterior** of θ given $\mathcal{D}$. Suppose we want to compare models, e.g., polynomial or neural networks, which model is better? We need to compare it by considering all possible parameter settings. The evidence is defined by:

$$P(\mathcal{D}|\mathcal{H}) = \int P(\mathcal{D}|\theta,\mathcal{H})P(\theta|\mathcal{H})d\theta$$

We will consider such Bayesian probabilistic models later.

3 Expectation, Variance

The expectation of a random variable x is:

$$E[x] = \sum_{x\in\mathcal{X}} xP(x)$$

or for continuous variables:

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

The variance of an random variable x:

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

Standard deviation σ is the square root of the variance. Standard deviation is the expectation of the distances from each data elements x to $E[x]$, or:

$$\sigma^2 = E[(x - E[x])^2] \tag{3}$$

Covariance of two random variables x and y is:

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

4 Information Theory

4.1 Entropy

Given two unrelated (independent) events x and y. The information we obtained from both events should be

$$h(x, y) = h(x) + h(y) \tag{4}$$

where $h(x)$ is an arbitrary measure for information. $h(x)$ should be a function of the probability distribution of x (i.e. $h(x) = f(p(x))$ and $f(\cdot)$ should be a function satisfying the commons sense that, an event is more rare, more information it could provide us. To satisfy the above two constraints. (Shannon 1948) employed negative logarithm as the measure for information content.

$$h(x) = -\log P(x) = \log \frac{1}{P(x)} \tag{5}$$

The **entropy** is an expectation of information measure $h(x)$.

$$H(x) = \int p(x) \log \frac{1}{p(x)} dx \tag{6}$$

For a discrete random variable x is defined as follows:

$$H(x) = \sum_{x \in \mathcal{X}} P(x) \log \frac{1}{P(x)} \tag{7}$$

The **joint entropy** of two random discrete variables x and y is defined as:

$$H(x, y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(x, y) \log \frac{1}{P(x, y)} \tag{8}$$

The **conditional entropy** of a discrete random variable x given the value of a discrete random variable $y = y'$ is defined as:

$$H(x|y = y') = \sum_{x \in \mathcal{X}} P(x|y = y') \log \frac{1}{P(x|y = y')} \tag{9}$$

The (expected) conditional entropy of x given y, or the **equivocation** of x about y is then given by:

$$H(x|y) = E[H(x|y)] = \sum_{y \in \mathcal{Y}} P(y) H(x|y) = \sum_{y \in \mathcal{Y}} P(y) \sum_{x \in \mathcal{X}} P(x|y) \log \frac{1}{P(x|y)} \tag{10}$$

where $P(x|y) = \frac{P(x,y)}{P(y)}$. We can simplify the above equation as:

$$H(x|y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(x,y) \log \frac{P(y)}{P(x,y)} \tag{11}$$

A basic property of the conditional entropy or equivocation is that:

$$H(x,y) = H(x|y) + H(y) \tag{12}$$

This equation can be proved as follows:

$$H(x|y) + H(y) = \sum_{x,y} P(x,y) \log \frac{P(y)}{P(x,y)} + \sum_{y} P(y) \log \frac{1}{P(y)} \tag{13}$$

Because the marginal probability $P(y) = \sum_x P(x,y)$, we can replace $P(y)$ in equation 13 and obtain:

$$\begin{aligned} H(x|y) + H(y) &= \sum_{x,y} P(x,y) \log \frac{P(y)}{P(x,y)} + \sum_{y} \sum_{x} P(x,y) \log \frac{1}{P(y)} && (14) \\ &= \sum_{x,y} P(x,y) \left[\log \frac{P(y)}{P(x,y)} + \log \frac{1}{P(y)} \right] && (15) \\ &= \sum_{x,y} P(x,y) \log \frac{1}{P(x,y)} = H(x,y) && (16) \end{aligned}$$

The **mutual information**, or transinformation is a measure of how much information can be obtained about one random variable by observing another. The mutual information of x relative to y is given by:

$$I(x;y) = \sum_{x,y} P(y) P(x|y) \log \frac{P(x|y)}{P(x)} \tag{17}$$

4.2 Kullback-Leibler Divergance and Gibb's Inequality

The Kullback-Leibler (KL) divergence or distance, or **relative entropy**, is a quantity which measures the difference between two probability distributions:

$$KL(p(x)||q(x)) = \int_x p(x) \log \frac{p(x)}{q(x)} dx \tag{18}$$

In Bayesian statistics the KL divergence can be used as a measure of the "distance" between the prior distribution and the posterior distribution.

There are also some interesting properties:

$$KL(p||q) + H(p) = H(p,q)$$

where

$$H(p,q) = E_p \left[\log \frac{1}{q(x)} \right] = \sum_x p(x) \log \frac{1}{q(x)}$$

is referred to as **cross entropy**. The relation between mutual information and KL divergence is as follows:

$$I(x;y) = KL(p(x,y)||p(x)p(y)) \tag{19}$$

The proof is as follows:

$$\begin{aligned}
I(x;y) &= \sum_{x,y} P(y)P(x|y)\log\frac{P(x|y)}{P(x)} && (20)\\
&= \sum_{x,y} P(x,y)\log\frac{P(x|y)P(y)}{P(x)P(y)} && (21)\\
&= \sum_{x,y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} && (22)\\
&= KL(P(x,y)||P(x)P(y)) && (23)
\end{aligned}$$

Given two distributions $p(x)$ and $q(x)$, their KL-divergence satisfy: $KL(p||q) \geq 0$. This is also called **Gibb's inequality**. To prove Gibb's inequality, we first need to know **Jensen's inequality**: given $f(x)$ is a convex function (see eq. 40) and x is a random variable:

$$E[f(x)] \geq f(E[x]) \tag{24}$$

In the following proof, we will use a convex function $f(x) = \log\frac{1}{x}$ and Jensen's inequality.

$$\begin{aligned}
KL(p||q) &= \int p(x)\log\frac{p(x)}{q(x)} && (25)\\
&= \int p(x)\log\frac{1}{q(x)/p(x)} && (26)\\
&= E\left[\log\frac{1}{q(x)/p(x)}\right] && (27)\\
&\geq \log\frac{1}{E[q(x)/p(x)]} && (28)\\
&= \log\frac{1}{\int p(x)\frac{q(x)}{p(x)}dx} && (29)\\
&= \log\frac{1}{1} = 0 && (30)
\end{aligned}$$

5 Probability Distributions

Bernoulli Distribution Given a discrete random variable x:

$$x \in \{0,1\} \quad P(x=1) = p \quad P(x=0) = 1-p$$

where p is the only parameter for the Bernoulli distribution whose **probability density function** (PDF) is described as the following:

$$f(x) = p^x(1-p)^{1-x}$$

Beta Distribution is a continuous probability distribution with probability density function (pdf) defined on the interval $[0,1]$:

$$f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \tag{31}$$

where the parameters α and β must be greater than zero and $B(\cdot)$ is the beta function defined by:

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt \tag{32}$$

Binomial Distribution is a discrete distribution and the PDF is:

$$f(x; p, n) = \binom{n}{x} p^x (1-p)^{n-x} \tag{33}$$

Gaussian Distribution is an extremely important proability distribution is statistics and its PDF is:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{34}$$

where $E(x) = \mu$ and $Var(x) = \sigma^2$. For multivariate Gaussian $\mathbf{x} \in \mathcal{R}^d$:

$$f(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right) \tag{35}$$

where $\mu \in \mathcal{R}^d$ is the d-dimensional mean and Σ is the d-by-d covariance matrix.

Dirichlet Distribution (after Johann Peter Gustav Lejeune Dirichlet) is a continuous multivariate probability distribution. The Dirichlet distribution is the multivariate generalization of the beta distribution (eq. 31). The PDE for the Dirichlet distribution of order K is a function of a K-dimensional vector $\mathbf{x} = \{x_1, \ldots, x_K\}$:

$$f(\mathbf{x}; \alpha) = Dir(\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{K} x_j^{\alpha_j - 1} \tag{36}$$

where $\sum_{i=1}^{K} x_i = 1$ and $\Gamma(\cdot)$ is Gamma function which defined by:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \tag{37}$$

Gamma function is an extension of the factorial function to the complex numbers.

$$\Gamma(t) = (t-1)\Gamma(t-1)$$

For integers:

$$\Gamma(t) = (t-1)!$$

The first term of eq. 36 is a normalizing constant: which is a multi-nomial beta function expressed in terms of the gamma function:

$$\frac{\prod_j \Gamma(\alpha_j)}{\Gamma(\sum_j \alpha_j)} = B(\alpha) \tag{38}$$

Hence, eq. 36 can be re-written as:

$$f(\mathbf{x}; \alpha) = \frac{1}{B(\alpha)} \prod_{j=1}^{K} x_j^{\alpha_j - 1} \tag{39}$$

6 Convex Optimization

A function $f(x)$ is convext over $[a, b]$ if for all $x_1, x_2 \in [a, b]$ and $0 \leq \lambda \leq 1$:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \tag{40}$$

If we denote $x' = \lambda x_1 + (1-\lambda)x_2$ where x' is actually a value in between of $[x_1, x_2]$ and $g(x) = \lambda f(x_1) + (1-\lambda) f(x_2)$. Fig. 1 gives a convex function where $f(x') \leq g(x)$. The following are some commonly used convex functions:

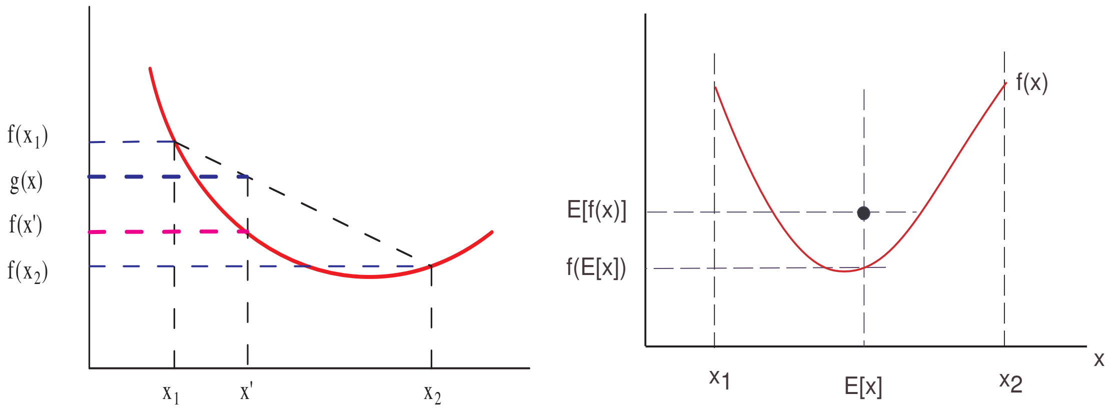

Figure 1: left: A convex function. Right: an schematic illustration of Jensen's inequality

- affine: $y = ax + b$
- exponential: $y = e^{ax}$
- negative logarithm: $y = -\log x$
- norm: $y = ||x||^2$
- quadratic: $y = ax^2 + bx + c$

Jensen's inequality Give a convex function.

$$E[f(x)] \geq f(E[x]) \tag{41}$$

This fact can be easily seen from the right-hand side figure of fig. 1. We can imagine the function is a rope with uniform mass. The weighting center is at $(E[x], E[f(x)])$. Since the function is convex, $f(E[x])$ is lower than $E[f(x)]$.

References

C. J. C. Burges (2004), Some notes on applied mathematics for machine learning, Bousquet et al. (Eds.), *Machine Learning 2003*, LNAI 3176, pp. 21-40. Springer-Verlag.

D. J. C. Mackay (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.

C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, July and October.

J. Wu (2005), Some properties of the normal distribution, Tutorial, source from: *http://www.cc.gatech.edu/∼wujx/paper/Gaussian.pdf*, Georgia Institute of Technology.