Probability

Notes on Spoken Language Processing

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Probability Theory

- Spoken language is random in nature.
- The best language to describe random processes is the probability.
- Make sure you know
  - random experiment
  - sample space
  - event

Conditional Probability

- The probability that event A occurs can be estimated by the relative frequency

$$p(A) = \frac{N_A}{N_S}$$

- The joint probability of events A and B is

$$p(A,B) = \frac{N_{AB}}{N_S}$$

- The conditional probability of event A given event B occurs is

$$p(A|B) = \frac{N_{AB}}{N_B} = \frac{p(A,B)}{p(B)}$$

Chain Rule

- The chain rule for probability is

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

- More generally,

$$p(A_1, \ldots, A_n) = p(A_n|A_1, \ldots, A_{n-1}) \ldots p(A_2|A_1)p(A_1)$$

Independence

- The condition that event B happens makes the probability of event A different. So $p(A|B)$ is generally different from $p(A)$.
- When $p(A|B) = p(A)$, we say that event A and event B are independent. It can be shown that

$$p(A, B) = p(A)p(B).$$

- We say that the probability factorizes.

Partition

- A partition of a set T is a set of disjoint sets whose union is T. The probability of a set T is the sum of probabilities of the sets in a partition of T.
- Let $A_1, \ldots, A_n$ be a partition of sample space S and B is any event. Then $B \cap A_1, \ldots, B \cap A_n$ is a partition of B.

$$p(B) = \sum_i p(A_i, B) = \sum_i (A_i) p(B|A_i).$$

The Bayes Rule

- From the conditional probability and the partition, we have

$$p(A_i|B) = \frac{p(A_i, B)}{p(B)} = \frac{p(A_i)p(B|A_i)}{\sum_k p(A_k)p(B|A_k)}$$

- The above is called the Bayes rule: one can see the posterior probability can be obtained from the prior probability and the conditional probability.

Random Variables

- A random variable X is a function $X(s)$ that maps an outcome s of a random experiment to a real number.
- $X = x$ defines an event that $\{s|X(s) = x\}$.
- The probability can be zero or non-zero depending on the nature of X, and the set of values it can take.

Discrete Random Variables

- X is said to be discrete if X can only takes discrete values.
- The function $p_X(x) = p(X = x)$ is the probability of event $\{s|X(s) = x\}$.
- $p_X(x)$ is a.k.a. the probability mass function (pmf).

Continuous Random Variables

- A random variable, say X, may take a continuum of values. It is said to be continuous.
- For an interval A, the probability for a continuous r.v. X to take a value in A can be written as

$$p(X \in A) = \int_A f_X(x)dx.$$

- In particular,

$$p(x < X \leqslant x + dx) = f_X(x)dx.$$

- $f_X(x)$ is called the probability density function (pdf) of X.

Distribution Functions

- Both discrete and continuous random variables can be characterized by distribution function defined by

$$F_X(x) = Pr(X \leq x).$$

- If X is continuous, the derivative of $F_X(x)$ is $f_X(x)$.
- If X is discrete, $F_X(x)$ is the sum of the probability masses less or equal to x,

$$F_X(x) = \sum_{x_i \leq x} Pr(X = x_i)$$

Joint Probability

- Two random variables may be related and we may want to describe them together.
- We define the joint distribution function for X and Y by

$$F(x, y) = Pr(X \leq x, Y \leq y).$$

- If X and Y are discrete, we can define the joint probability mass function by

$$p(x, y) = Pr(X = x, Y = y).$$

Joint Density Function

- If both X and Y are continuous, we can define the joint density function $f(x, y)$ by the distribution function

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv.$$

- That is,

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

- Note that $F(x, y)$ is non-decreasing in "upper-right" directions and $f(x, y)$ is non-negative everywhere.

Conditional Probability: Discrete

- For two events A and B, the conditional probability that A occurs given that B occurs is defined by

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}.$$

- For two discrete RVs, X and Y, the conditional probability of $\{Y = y\}$ given $\{X = x\}$ is

$$p_{Y|X}(y|x) \triangleq Pr(Y = y|X = x) = \frac{Pr(X = x, Y = y)}{Pr(X = x)} = \frac{p(x, y)}{p(x)}.$$

Conditional Probability: Continuous

- Suppose that X and Y are continuous.
- How can we define the conditional probability?
- We can define a density function via infinitesimal probability masses as follows

$$\frac{Pr(y < Y \leq y + dy,\ x < X \leq x + dx)}{Pr(x < X \leq x + dx)} = \frac{f(x,y)dxdy}{f(x)dx}$$
$$= \frac{f(x,y)dy}{f(x)} = f(y|x)dy.$$

- It follows that

$$f(y|x) = \frac{f(x,y)}{f(x)}.$$

Conditional Probability: Mixed case

- The last case we will discuss is the mixed case: Y is discrete and X is continuous.
- Again, we can define a mass function $p(y|x)$ via infinitesimal probability masses as follows

$$\frac{Pr(Y = y,\ x < X \leq x + dx)}{Pr(x < X \leq x + dx)} = \frac{f(x,y)dx}{\sum_{y'} f(x,y')dx}$$

$$\Rightarrow \quad p(y|x) = \frac{f(x,y)}{\sum_{y'} f(x,y')} = \frac{p(y)f(x|y)}{\sum_{y'} p(y')f(x|y')}.$$

- In pattern recognition we often have continuous features and discrete class labels. The conditional probability of a class given features is well-defined.

Probability Equality: Discrete

- The marginalization, chain rule and Bayes rule apply to discrete random variables,

$$p_X(x_i) = \sum_j p_{XY}(x_i, y_j)$$

$$p_{XY}(x_i, y_j) = p_X(x_i) p_{Y|X}(y_j|x_i) = p_Y(y_j) p_{X|Y}(x_i|y_j)$$

$$\begin{aligned} p_{X|Y}(x_i|y_j) &= \frac{p_{XY}(x_i, y_j)}{p_Y(y_j)} \\ &= \frac{p_{Y|X}(y_j|x_i) p_X(x_i)}{\sum_k p_{Y|X}(y_j|x_k) p_X(x_k)} \end{aligned}$$

Independence of Random Variables

- Two random variables X and Y are independent if

$$p_{X|Y}(x|y) = p_X(x) \;\; \forall\, x, y.$$

- Independence is also characterized by factorization

$$p_{XY}(x, y) = p_X(x) p_Y(y).$$

Probability Equality: Continuous

- The marginalization, chain rule and Bayes rule apply to continuous random variables as well, but in the following form

$$f_X(x) = \int f_{XY}(x,y)dy$$

$$f_{XY}(x,y) = f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y)$$

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{XY}(x,y)}{f_Y(y)} \\ &= \frac{f_{Y|X}(y|x)f_X(x)}{\int f_{Y|X}(y|x)f_X(x)dx} \end{aligned}$$

Mean and Variance

- The expectation value of a function of random variable $g(X)$ is defined by

$$E(g(X)) = \begin{cases} \sum_i p_X(x_i)g(x_i), & X \text{ discrete} \\ \int_x f_X(x)g(x), & X \text{ continuous} \end{cases}$$

- The mean of X is the expectation value of X

$$\mu_X = E(X).$$

- The variance of X is defined by

$$Var(X) = \sigma_X^2 = E[(X - \mu_X)^2].$$

Conditional Expectation

- Suppose X and Y are random variables. The conditional expectation of Y given $X = x$ is

$$E(Y|X = x) = \sum_y y \, p_{Y|X}(y|x).$$

- This can be seen as a function of random variable X.
- The expectation of this function (w. r. t. X) is

$$\begin{aligned} E_X[E(Y|X)] &= \sum_x p(x) \sum_y y \, p_{Y|X}(y|x) \\ &= \sum_{x,y} p(x,y) \, y = E_{XY}(Y) = E_Y(Y) \end{aligned}$$

Covariance

- Suppose X and Y are random variables.
- The covariance of X and Y is defined by

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

- The correlation coefficent of X and Y, denoted by ρ_{XY}, is defined by

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Theorems

- For any random variables X and Y

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

- For independent random variables X and Y

$$\rho_{XY} = 0 = Cov(X,Y).$$

- For any random variables $X_1, \ldots, X_n$

$$Var(\sum_i X_i) = \sum_i Var(X_i) + \sum_i \sum_{j \neq i} Cov(X_i, X_j)$$

Random Vectors

- We denote a random vector by

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)^T.$$

- Each component in $\mathbf{X}$ is a scalar random variable. They may or may not be independent.

Mean Vector and Covariance Matrix

- The mean vector is defined by

$$\mu_X = E(\mathbf{X}) = (E(X_1), \ldots, E(X_n))^T$$

- The covariance matrix is defined by

$$\Sigma_X = Cov(\mathbf{X}) = E[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^T]$$

$$= \begin{bmatrix} Cov(X_1, X_1) & \ldots & Cov(X_1, X_n) \\ \vdots & \ddots & \vdots \\ Cov(X_n, X_1) & \ldots & Cov(X_n, X_n) \end{bmatrix}$$

Linear Transformation

- For a linear transformation of a random vector

$$Y = AX + B,$$

we have

$$\mu_Y = A\mu_X + B,$$

and

$$\Sigma_Y = A\Sigma_X A^T.$$

Common Distributions

- uniform
- binomial
- geometric
- multinomial
- Poisson
- Gamma
- Gaussian

Uniform Distributions

- Everything is equally likely.
  - discrete case

$$p_X(x_i) = \texttt{const}$$

  - continuous case

$$f(x) = \texttt{const}$$

- This is the distribution of a random variable whose value is the most difficult to predict.

Binomial Distributions

- Suppose in a toss of a coin, the outcome is a head with probability p.
- Let X be the number of heads in n tosses. Then

$$p_X(i) = \binom{n}{i} p^i (1-p)^{n-i}.$$

- The above is called a binomial distribution $B(n, p)$. It can be shown that

$$\mu_X = np,\ \sigma_X^2 = np(1-p).$$

Geometric Distributions

- Consider the previous coin. The number of tosses G until the first tail shows up is a random variable.
- The distribution of G is

$$p_G(j) = p^{j-1}(1-p).$$

- The above is called a geometric distribution. It can be shown that

$$\mu_G = \frac{1}{1-p},\ \sigma_G^2 = \frac{1}{(1-p)^2}.$$

- The duration of a hidden Markov model state follows this distribution.

Multinomial Distributions

- From an urn of k different colors of balls we pick sequentially n balls with replacement.
- Let p_i be the probability that a ball of color i is picked.
- Let X_i be the number of picks of a ball of color i. The distribution for the random vector $(X_1, \ldots, X_k)$ is called a multinomial distribution.

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}, \ \sum_k x_k = n, x_k \geq 0.$$

$$\mu_{X_i} = np_i, \sigma_{X_i}^2 = np_i(1-p_i), Cov(X_i, X_j) = -np_ip_j.$$

Poisson Distributions

- The Poisson distribution for a non-negative, integer-valued random variable with parameter λ is

$$p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

- It can be shown that

$$\mu_X = \lambda, \sigma_X^2 = \lambda.$$

- Poisson distribution is often used in queuing theory, to characterize the total number of arrivals (or departures) in a time unit.

Gamma Distributions

- A non-negative continuous random variable X has a Gamma distribution with parameters $\alpha > 0, \beta > 0$ if

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

where $\Gamma(\cdot)$ is the Gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

- There are some interesting properties regarding Gamma distributions which we will mention when we need to.

Gaussian (Normal) Distributions

- A continuous random variable X is siad to have a Gaussian distribution with parameter μ, σ if

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Gaussian distribution is a.k.a. the normal distribution. The above function is also denoted by

$$N(x; \mu, \sigma^2).$$

- Gaussian distributions approximate many uni-modal distributions.

Standard Gaussian

- The standard Gaussian distribution refers to $N(x; 0, 1)$.
- It is a Gaussian distribution with zero mean and unit variance.
- A random variable X with Gaussian distribution can be "standardized"or "normalized" by a transform

$$Z = \frac{X - \mu}{\sigma} \Rightarrow p_Z(z) = N(z; 0, 1).$$

Central Limit Theorem

- Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with mean μ and σ^2. Define S_n and $\bar{X}_n$ by

$$S_n = \sum_{i=1}^{n} X_i = n\bar{X}_n.$$

- (theorem) S_n approaches a Gaussian with mean $n\mu$ and variance $n\sigma^2$.
- Sample mean $\bar{X}_n$ approaches a Gaussian with mean μ and variance σ^2/n. I.e.

$$Y = \lim_{n\to\infty} \frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(y; 0, 1)$$

Multi-variate Gaussian Distributions

- Let $\mathbf{X} = (X_1, \ldots, X_n)'$ be a random vector.
- $\mathbf{X}$ is said to have a multi-variate Gaussian distribution if

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)}$$

- Note it reduces to the uni-variate Gaussian if $n = 1$.
- Σ is diagonal if the X_i's are mutually independent.

Gaussian Mixture

- A random vector $\mathbf{X}$ is said to have a K-component Gaussian mixture distribution if

$$f_{\mathbf{X}}(\mathbf{x}) = \sum_{k=1}^{K} c_k N(\mathbf{x}; \mu_{\mathbf{k}}, \mathbf{\Sigma_k}),$$

where

$$\sum_{k=1}^{K} c_k = 1; \quad c_k \geq 0.$$

- A Gaussian mixture distribution can approximate any distribution when K is large enough.

χ^2 Distributions

- A χ^2 distribution with n degrees of freedom is a special case of Gamma distribution where $\alpha = \frac{n}{2}, \beta = \frac{1}{2}$.

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

- (theorem) If $X_1, \ldots, X_n$ are i.i.d. standard Gaussian, then

$$Z = X_1^2 + \cdots + X_n^2$$

has a χ^2-distribution with n degrees of freedom.

Log-Normal Distributions

- Y is log-normal if $\log Y$ is normal.
- That is, let X be normal. Then $Y = e^X$ is log-normal.
- The distribution function for Y can be found via

$$f_Y(y)dy = f_X(x)dx.$$

- From the above it follows

$$f_Y(y) = \frac{1}{y} N(\log y; \mu_X, \sigma_X^2)$$

Estimation

- Often the true distribution is unknown.
- We have samples from an unknown distribution. We may be able to learn something about the unknown distribution from these samples.
- Let Φ denote a set of parameters. The problem is to estimate Φ from data.
- This is called parameter estimation.

Estimator and Estimate

- An estimator is a function that specifies parameter value for all possible samples.
- It is itself a random variable, which can be denoted by $\theta(X_1, \ldots, X_n)$.
- Note the mean and variance of an estimator are well-defined.
- An estimate is a specific value of the estimator with specific sample values $\theta(x_1, \ldots, x_n)$.

MMSE Estimation

- The minimum mean squared error (MMSE) estimator is the function $\hat{Y} = g(X)$ such that the expected squared error is minimized.
- Let $g(X)$ be parameterized by $g(X, \Phi)$, then

$$\Phi_{\text{MMSE}} = \arg\min_{\Phi} E[(g(X, \Phi) - Y)^2]$$

LSE Estimation

- Often the distribution is unknown but we have samples $\{(x_i, y_i), i = 1, 2, \ldots, n\}$.
- The least squared error (LSE) estimation is applied with unknown distribution

$$\Phi_{\text{LSE}} = \arg\min_{\Phi} \sum_{i=1}^{n} (g(x_i, \Phi) - y_i)^2$$

Constant Functions

- The simplest family of functions is the constant function

$$\hat{Y} = g(X) = c,$$

where c is the parameter to be decided.

- Minimizing $E[(\hat{Y} - Y)^2] = E[(c - Y)^2]$ over c yields

$$c_{\text{MMSE}} = \mu_Y .$$

- The LSE estimate can be shown to be

$$c_{\text{LSE}} = \frac{1}{n} \sum y_i$$

Linear Functions

- The family of linear functions is

$$\hat{Y} = g(X) = aX + b,$$

where a, b are to be decided.

- Minimizing $E[(\hat{Y} - Y)^2] = E[(aX + b - Y)^2]$ over a, b yields

$$a = \frac{Cov(X,Y)}{Var(X)} = \rho_{XY}\frac{\sigma_Y}{\sigma_X};$$

$$b = \mu_Y - a\mu_X.$$

Vectors

- Suppose Y is a scalar and X is a d-dim vector. We want to find $(b = a_0, a)$ that minimizes the LSE
- Let $\mathbf{y}$ be an $n \times 1$ column vector consisting of y_i, $\mathbf{X}$ be an $n \times (d+1)$ matrix consisting of $(1, x_i)$'s as row vectors, and $\mathbf{a} = (a_0, a)'$, then

$$e(\mathbf{a}) = ||\hat{\mathbf{y}} - \mathbf{y}||^2 = ||\mathbf{X}\mathbf{a} - \mathbf{y}||^2$$

- $\mathbf{a}$ can be solved by the normal equation

$$\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{a}_{\text{LSE}} = \mathbf{X}^{\mathbf{T}}\mathbf{y} \Rightarrow \mathbf{a}_{\text{LSE}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{y}.$$

Maximum Likelihood Estimation

- We determine the parameter values to maximize the probability of the data (called data likelihood).
- Let Φ be the parameter set and $\mathbf{x} = (x_1, \ldots, x_n)$ be the data, then

$$\Phi_{\text{MLE}} = \arg\max_{\Phi} p(\mathbf{x}|\Phi) = \arg\max_{\Phi} \prod_i p(x_i|\Phi).$$

- We can also work on the log likelihood

$$\Phi_{\text{MLE}} = \arg\max_{\Phi} \log p(\mathbf{x}|\Phi) = \arg\max_{\Phi} \sum_i \log p(x_i|\Phi).$$

Properties

- Φ_{MLE} is a random variable whose distribution is decided by the distribution of X.
- As the number of samples grows, Φ_{MLE} has a Gaussian distribution with a mean $\tilde{\Phi}$, the true parameter, and a variance inversely proportional to n. So

$$\lim_{n \to \infty} \Phi_{\text{MLE}} = \tilde{\Phi}.$$

- Φ_{MLE} is said to be a *consistent* estimator.

Bayesian Estimation

- Bayesian estimation treats a parameter Φ as a random variable.
- Φ has a prior distribution $p(\Phi)$ that is turned into a posterior distribution $p(\Phi|\mathbf{x})$ after samples $\mathbf{x}$ are observed.
- According to Bayes' rule

$$p(\Phi|\mathbf{x}) = \frac{p(\mathbf{x}|\Phi)p(\Phi)}{p(\mathbf{x})} \propto p(\mathbf{x}|\Phi)p(\Phi)$$

General Bayesian Estimation

- We define a loss (risk) function $R(\Phi, \bar{\Phi})$.
- The expected risk is minimized,

$$E[R(\Phi, \bar{\Phi})] = \int R(\Phi, \bar{\Phi}) p(\Phi) d\Phi.$$

- When $\mathbf{x}$ is observed, $p(\Phi)$ is replaced by $p(\Phi|\mathbf{x})$, and

$$\theta_{\text{Bayes}}(\mathbf{x}) = \arg\min_{\theta} E[R(\Phi, \theta(\mathbf{x})) \mid \mathbf{x}].$$

- The solution depends on the risk function, as well as the distribution of Φ.

Conjugate Priors

- A conjugate prior is a probability function such that $p(\Phi)$ and $p(\Phi|\mathbf{x})$ belong to the same probability family.
- For mathematical tractability, conjugate priors are often used in Bayesian estimation.
- A common example is the Gaussian conjugate prior, where the prior and the posterior distributions of mean, and the conditional distribution of data, are all Gaussians.

MAP Estimation

- Maximum a posteriori estimation chooses an estimate that maximizes the posterior distribution.

$$\Phi_{\mathrm{MAP}} = \arg\max_{\Phi} p(\Phi|\mathbf{x}) = \arg\max_{\Phi} p(\mathbf{x}|\Phi)p(\Phi)$$

- Apparently, if the prior term is a constant, then MAP estimator is the same as the ML estimator.
- The prior function can be seen as knowledge about the parameter Φ.
- When the size of training data is limited, such information may be valuable.

Information Theory

- originally developed by Shannon in his analysis of reliable transmission of data over communication channels.
- deal with problems of encoding, transmission, decoding

Information

- The quantity of information of an event can be measured by

$$I(x_i) = \log \frac{1}{p(x_i)}$$

- From this definition, the more unlikely an event occurs, the more information is provided when it does occur.
- The average information of a random source is called its entropy, which we define next.

Entropy

- The most fundamental concept of information theory is the entropy.
- The entropy of a random variable X is defined by

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)}$$

- The entropy is non-negative.
  - It is zero when the random variable value is “certain”.
  - A uniform distribution has the maximum entropy.

Perplexity

- The perplexity of a source X is defined by

$$PP(X) = 2^{H(X)}.$$

- Perplexity has an interpretation in natural language processing: it is the branching factor of a sentence.

Joint and Conditional Entropy

- For two random variables X and Y, the joint entropy is defined by

$$H(X,Y) = \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)}.$$

- The conditional entropy is defined by

$$\begin{aligned} H(X|Y) &= \sum_{y} p(y) H(X|Y=y) \\ &= \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)}. \end{aligned}$$

Mutual Information

- The mutual information of X and Y is defined by

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

- It follows from definition of entropy and mutual information that

$$I(X;Y) = H(X) - H(X|Y).$$

- Some training methods use mutual information as the objective function.

Source Coding Theorem

- To encode a random information source with zero probability of error for decoding, the number of bits per symbol must be at least $H(X)$.

Channel Coding Theorem

- If the bit rate R is not greater than the channel capacity C of a communication system, then there exists an error-free transmission method.