CSCI567 Machine Learning (Fall 2016)

Dr. Yan Liu

yanliu.cs@usc.edu

August 28, 2016

Outline

Registration and Contact Information

Registration

- D-clearance has been given to all students who passed the entrance exam and on borderline cases.
- Please register as soon as possible so that we can have the final counts.

Contact Information

- Please use the general email: CSCI567.usc@gmail.com for all future email communications.
- The TAs will stop responding to individual requests from now on.

Forums and Discussion Sessions

Forums on Blackboard

- Please ask all your technical questions in the appropriate discussion forums.
- The TAs and the instructor will answer the questions within 24 hours.
- If you haven't received your answer within 24 hours, please send an email to CSCI567.usc@gmail.com. It will be directed to my attention.

Discussion Sessions

- The TAs will send out the contents of the discussion session early in the week.
- There will be only one discussion session.

Outline

How to grasp machine learning well

Three pillars to machine learning[1]

- Probability, Statistics and Information Theory
- Linear Algebra and Matrix Analysis
- Optimization

Resources to study them

- Suggested Reading:
  - All of Statistics Page 21-89
  - Murphy's textbook
- URL pointers on the syllabus
- Wikipedia (some information might not be 100% accurate, though)

Quote from Prof. Michael I. Jordan

Outline

Probability: basic definitions

Sample Space: a set of all possible outcomes or realizations of some random trial.
Example: Toss a coin twice; the sample space is
$\Omega = \{HH, HT, TH, TT\}$.

Event: A subset of sample space
Example: the event that at least one toss is a head is
$A = \{HH, HT, TH\}$.

Probability: We assign a real number $P(A)$ to each event A, called the probability of A.

Probability Axioms: The probability P must satisfy three axioms:

1. $P(A) \geq 0$ for every A;
2. $P(\Omega) = 1$;
3. If $A_1, A_2, \ldots$ are disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Random Variables

Definition: A random variable is a measurable function that maps from a probability space to a measurable space, i.e. $X : \Omega \to R$, that assigns a real number $X(\omega)$ to each outcome ω.

Example: if $\Omega = \{(x, y) : x^2 + y^2 \leq 1\}$ and our outcomes are samples (x, y) from the unit disk, then these are some examples of random variables: $X(\omega) = x$, $Y(\omega) = y$, $Z(\omega) = x + y$.

Data and Statistics The data are specific realizations of random variables; A statistics is just any function of the data or random variables.

Distribution Function

Definition: Suppose X is a random variable, x is a specific value that it can takes,
Cumulative distribution function (CDF) is the function $F : R \rightarrow [0, 1]$, where $F(x) = P(X \leq x)$.

If X is discrete $\Rightarrow$ *probability mass function*: $f(x) = P(X = x)$.
If X is continuous $\Rightarrow$ *probability density function* for X if there exists a function f such that $f(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f(x)dx = 1$ and for every $a \leq b$,

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

If $F(x)$ is differentiable everywhere, $f(x) = F'(x)$.

Expectation

Expected Values

- Discrete random variable X, $E[g(X)] = \sum_{x \in \mathcal{X}} g(x)f(x)$;
- Continuous random variable X, $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)$

Mean and Variance $\mu = E[X]$ is the mean; $var[X] = E[(X - \mu)^2]$ is the variance.
We also have $var[X] = E[X^2] - \mu^2$.

Common Distributions

Discrete variable	Probability function	Mean	Variance
Uniform $X \sim U[1,\dots,N]$	$1/N$	$\frac{N+1}{2}$	
Binomial $X \sim Bin(n,p)$	$\binom{n}{x} p^x (1-p)^{(n-x)}$	np	
Geometric $X \sim Geom(p)$	$(1-p)^{x-1}p$	$1/p$	
Poisson $X \sim Poisson(\lambda)$	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	
Continuous variable	Probability density function	Mean	Variance
Uniform $X \sim U(a,b)$	1/ (b-a)	(a + b)/2	
Gaussian $X \sim N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$	μ	
Gamma $X \sim \Gamma(\alpha,\beta)$ $(x \geq 0)$	$\frac{1}{\Gamma(\alpha)\beta^a} x^{a-1} e^{-x/\beta}$	$\alpha\beta$	
Exponential $X \sim exponen(\beta)$	$\frac{1}{\beta} e^{-\frac{x}{\beta}}$	β	

Multivariate Distributions

Definition:

$$F_{X,Y}(x,y) := P(X \leq x, Y \leq y),$$

and

$$f_{X,Y}(x,y) := \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y},$$

Marginal Distribution of X (Discrete case):

$$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f_{X,Y}(x,y)$$

or $f_X(x) = \int_y f_{X,Y}(x,y)dy$ for continuous variable.
Conditional probability of X given $Y = y$ is

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

Transformation of Random Variables

Let $\mathbb{X} = (X_1, \ldots, X_k)$ be a k-dimensional random variable with pdf $f_{\mathbf{X}}(\mathbf{x})$. define a differentiable transformation of $\mathbf{X}$ into $\mathbf{Y}$ using g, such that

$$g(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_k(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} = \mathbf{y}$$

with the inverse $h(\mathbf{y}) = \mathbf{x}$.
The pdf of $\mathbf{Y}$ is $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{x}}(h(\mathbf{y}))|J(\mathbf{x}, \mathbf{y})|$, where

$$|J(\mathbf{x}, \mathbf{y})| = |\mathrm{det}(\begin{bmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_k}{\partial y_1} & \cdots & \frac{\partial h_k}{\partial y_k} \end{bmatrix})|$$

In-class Exercise

Suppose X is a random variable, following the *standard normal* distribution

$$X \sim N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Then, what is the distribution for $Y = X^2$?

Example

Suppose X is a random variable, following the *standard normal* distribution

$$X \sim N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Then, what is the distribution for $Y = X^2$?

- $X = h_1(Y) = \sqrt{Y}$ or $X = h_2(Y) = -\sqrt{Y}$
- We need to consider each branch, thus

$$f_Y(y) = f_X(h_1(y))|\frac{dh_1(y)}{dy}| + f_X(h_2(y))|\frac{dh_2(y)}{dy}| = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}$$

This distribution is called χ^2-distribution.

Bayes Rule

Law of total Probability: X takes values $x_1, \ldots, x_n$ and y is a value of Y, we have

$$f_Y(y) = \sum_j f_{Y|X}(y|x_j) f_X(x_j)$$

Bayes Rule:
(Simple Form)

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

(Discrete Random Variables)

$$f_{X|Y}(x_i|y) = \frac{f_{Y|X}(y|x_i) f_X(x_i)}{\sum_j f_{Y|X}(y|x_j) f_X(x_j)}$$

(Continuous Random Variables)

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{\int_x f_{Y|X}(y|x) f_X(x) dx}$$

Independence

Independent Variables X and Y are *independent* if and only if:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

or $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all values x and y.

IID variables: *Independent and identically distributed* (IID) random variables are drawn from the same distribution and are all mutually independent.

If $X_1, \dots, X_n$ are independent, we have

$$E[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} E[X_i], \quad var[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i^2 var[X_i]$$

Linearity of Expectation: Even if $X_1, \dots, X_n$ are not independent,

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i].$$

Correlation

Covariance

$$cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)],$$

Correlation coefficients

$$corr(X,Y) = Cov(X,Y)/\sigma_x\sigma_y$$

- Independence $\Rightarrow$ Uncorrelated $(corr(X,Y) = 0)$.

However, the reverse is generally not true.
The important special case: multi-variate Gaussian distribution.

Outline

Statistics

Suppose $X_1, \ldots, X_n$ are random variables:

Sample Mean:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Sample Variance:

$$S_{N-1}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2.$$

If X_i are iid:

$$\begin{aligned} E[\bar{X}] &= E[X_i] = \mu, \\ Var(\bar{X}) &= \sigma^2/N, \\ E[S_{N-1}^2] &= \sigma^2 \end{aligned}$$

Point Estimation

Definition The *point estimator* $\hat{\theta}_N$ is a function of samples $X_1, \ldots, X_N$ that approximates a parameter θ of the distribution of X_i.

Sample Bias: The bias of an estimator is

$$bias(\hat{\theta}_N) = E_\theta[\hat{\theta}_N] - \theta$$

An estimator is *unbiased estimator* if $E_\theta[\hat{\theta}_N] = \theta$

Standard error The standard deviation (i.e. the square-root of variance) of $\hat{\theta}_N$ is called the *standard error*

$$se(\hat{\theta}_N) = \sqrt{Var(\hat{\theta}_N)}.$$

Example

Suppose we have observed N realizations of the random variable X:

$$x_1, x_2, \cdots, x_N$$

Then,

- Sample mean $\bar{X} = \frac{1}{N}\sum_n x_n$ is an unbiased estimator of X's mean.
- Sample variance $S^2_{N-1} = \frac{1}{N-1}\sum_n (x_n - \bar{X})^2$ is an unbiased estimator of X's variance
- Sample variance $S^2_N = \frac{1}{N}\sum_n (x_n - \bar{X})^2$ is *not* an unbiased estimator of X's variance

Another example

Suppose we have observed N realizations of the random variable X:

$$x_1, x_2, \cdots, x_N$$

Moreover, suppose we know the true value of X's mean μ. Then,

- Sample variance $S_{N-1}^2 = \frac{1}{N-1}\sum_n (x_n - \mu)^2$ is *not* an unbiased estimator of X's variance
- Sample variance $S_N^2 = \frac{1}{N}\sum_n (x_n - \mu)^2$ is an unbiased estimator of X's variance

More example

Suppose we have observed N realizations of the random variable X:

$$x_1, x_2, \cdots, x_N$$

Then, in general, neither $\sqrt{S^2_{N-1}}$ nor $\sqrt{S^2_N}$ is an unbiased estimator for σ, i.e., the standard deviation of X.

Outline

Review on Information Theory

Suppose X can have one of the m values: $x_1, \ldots, x_m$. The probability distribution is $P(X = x_i) = p_i$.

Entropy is the smallest possible number of bits, on average, per symbol, needed to transmit a steam of symbols drawn from distribution of X.

$$H(X) = -\sum_{i=1}^{m} p_i \log p_i$$

- "High entropy" means X is from a uniform (boring) distribution;
- "Low entropy" means X is from varied (peaks and valleys) distribution.

Information Theory

Conditional Entropy is the remaining entropy of a random variable Y given that the value of another random variable X is known.

$$H(Y|X) = \sum_{i=1}^{m} p(X = x_i) H(Y|X = x_i) = -\sum_{i=i}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log p(y_j|x_i)$$

Mutual Information: if Y must be transmitted, how many bits on average would be saved if both ends of the line knew X?

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

Notice that $I(Y; X) = I(X; Y) = H(X) + H(Y) - H(X, Y)$
Kullback-Leibler divergence is a measure of distance between two distributions: a "true" distribution $p(X)$, and an arbitrary distribution $q(X)$.

$$\mathsf{KL}(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

We can write $I(X; Y) = KL(p(x, y)||p(x)p(y))$.

Outline

Optimization

Definition: Optimization refers to choosing the best element from some set of available alternatives. A general form is as follows:

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, i = 1, \dots, m \\ & h_i(x) = 0, i = 1, \dots, p. \end{aligned} \tag{1}$$

Difficulties:

1. Local or global optimimum?
2. Difficulty to find a feasible point,
3. Stopping criteria,
4. Poor convergence rate,
5. Numerical issues

Convex Optimization

Convex Functions: if for any two points x_1 and x_2 in its domain X and any $t \in [0, 1]$,

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2).$$

A function f is said to be *concave* if $-f$ is convex.
Convex Set a set S is convex if and only if for any $x_1, x_2 \in S$, $tx_1 + (1-t)x_2 \in S$ for any $t \in [0, 1]$,
Convex Optimization is minimization (maximization) of a convex (concave) function over a convex set.
Examples: Linear Programming (LP), Quadratic Programming (QP), and Semi-Definite Programming (SDP).
Popular convex optimization algorithms:

- Gradient descent
- Conjugate gradient
- Newton's method
- Quasi-Newton method
- Subgradient method