CS M146 - Week 1

Xinzhu Bei

xzbei@cs.ucla.edu

January 12, 2018

Overview

- Miscellaneous
  - Xinzhu Bei, xzbei@cs.ucla.edu
  - Discussion: Friday 2:00 - 3:50 pm, PUB AFF 1337
  - Office Hour: Monday 12-2 pm, Eng VI 386 (Tentative)
- Suggested Math Resources
  - Linear Algebra Review and Reference by Zico Kolter and Chuong Do: http://cs229.stanford.edu/section/cs229-linalg.pdf
  - Probability Theory Review by Arian Maleki and Tom Do: http://cs229.stanford.edu/section/cs229-prob.pdf
  - Convex Optimation Review by Zico Kolter and Honglak Lee: https://see.stanford.edu/materials/aimlcs229/cs229-cvxopt.pdf

Linear Algebra Review - Basic Notation

- By $A \in R^{m \times n}$ we denote a matrix with m rows and n columns

$$A = \begin{bmatrix} a_{11} a_{12} & \cdots a_{1n} \\ a_{21} a_{22} & \cdots a_{2n} \\ & \cdots \\ a_{m1} a_{m2} & \cdots a_{mn} \end{bmatrix}$$

- By $x \in R^n$, we denote a vector with n entries.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$$

Linear Algebra Review - Multiplication

- **Matrix Multiplication**: The product of two matrices $A \in R^{m \times n}$ and $B \in R^{n \times p}$ is the matrix

$$C = AB \in R^{m \times p}, \quad \text{where} C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

- **Vector-Vector Product**(sometimes called the **inner product** or dot product of the vectors): Given two vectors $x, y \in R^n$,

$$x^T y \in R = [x_1 x_2 \cdots x_n] \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i$$

Linear Algebra Review - Multiplication

- **Matrix-Vector Products**:

$$
\begin{aligned}
y = Ax &= \begin{bmatrix} -a_1^T & - \\ -a_2^T & - \\ \dots & \\ -a_m^T & - \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \dots \\ a_m^T x \end{bmatrix} \\
&= \begin{bmatrix} | & | & & | \\ a_1 & a_2 \cdots & & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = [a_1]x_1 + [a_2]x_2 + \cdots + [a_n]x_n
\end{aligned} \tag{1}
$$

Linear Algebra Review - The Inverse

- Example: consider the linear system of equations, $Ax = b$ where $A \in R^{n \times n}$, and $x, b \in R^n$. If A is invertible, then $x = A^{-1}b$.
- The **inverse** of a square matrix $A \in R^{n \times n}$ is denoted A^{-1}, and is the unique matrix such that $A^{-1}A = I = AA^{-1}$.
- A square matrix A has an inverse iff the determinant $|A| \neq 0$.
- In particular, we say that A is **invertible** or **non-singular** if A^{-1} exists and **non-invertible** or **singular** otherwise.

Linear Algebra Review - The Inverse

- Example: How to calculate inverse?

$$\left[\begin{array}{ccc|ccc} 1 & 3 & 3 & 1 & 0 & 0 \\ 1 & 4 & 3 & 0 & 1 & 0 \\ 1 & 3 & 4 & 0 & 0 & 1 \end{array}\right] \xrightarrow[-R_1+R_3]{-R_1+R_2} \left[\begin{array}{ccc|ccc} 1 & 3 & 3 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{-3R_2+R_1} \left[\begin{array}{ccc|ccc} 1 & 0 & 3 & 4 & -3 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{array}\right]$$

$$\xrightarrow{-3R_3+R_1} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 7 & -3 & -3 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{array}\right]$$

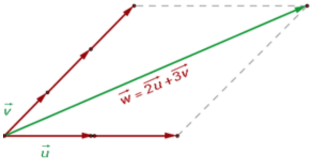
- Example: A general case of 2×2 matrix

$$\left[\begin{array}{cc|cc} a & b & 1 & 0 \\ c & d & 0 & 1 \end{array}\right]$$

- If

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for some scalar values $\alpha_1, \cdots, \alpha_{n-1} \in R$, then we say that the vectors $x_1, \cdots, x_n$ are **linearly dependent**; otherwise, the vectors are **linearly independent**.



- The **rank** of a matrix $A \in R^{m \times n}$ is the size of the largest subset of columns(rows) of A that constitute a linearly independent set.

Linear Algebra Review - Norm

A **norm** of a vector $\|x\|$ is informally a measure of the length of the vector.

- L2-norm: $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$, Note that $\|x\|_2^2 = x^T x$
- l1-norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- l∞-norm: $\|x\|_\infty = \max_i |x_i|$
- lp-norm: $\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{1/p}$

Linear Algebra Review - The Determinant

The **determinant** of a square matrix $A \in R^{n \times n}$, is a function det : $R^{n \times n} \rightarrow R$, and is denoted $|A|$ or $\det A$.

Geometric interpretation: given a matrix

$$\begin{bmatrix} -a_1^T- \\ \dots \\ -a_n^T- \end{bmatrix}$$

consider the set of points $S \subset R^n$

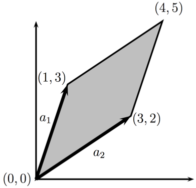
$$S = \{v \in R^n : v = \sum_{i=1}^{n} \alpha_i \alpha_i, \text{ where } 0 \le a_i \le 1, i = 1, \cdots, n\}$$

The absolute value of the determinant of A, it turns out, is a measure of the volume of the set S.

Linear Algebra Review - The Determinant

- Example: consider the 2×2 matrix

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}, \text{ where } a_1 = [1,3]^T; a_2 = [3,2]^T$$



the set S corresponds to the shaded region (i.e., the parallelogram).

- The general (recursive) formula for the determinant is:

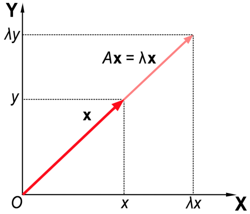
$$\begin{aligned} |A| &= \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |A_{\backslash i, \backslash j}| \text{ for any } j \in 1, \cdots, n \\ &= \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |A_{\backslash i, \backslash j}| \text{ for any } i \in 1, \cdots, n \end{aligned} \tag{2}$$

Linear Algebra Review - Eigenvalues and Eigenvectors

- Given a square matrix $A \in R^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an **eigenvalue** of A and $x \in \mathbb{C}^n$ is the corresponding **eigenvector** if

$$Ax = \lambda x, x \neq 0$$

  We assume that the eigenvector is normalized to have length 1.



- We can rewrite the equation above to state that (λ, x) is an eigenvalue-eigenvector pair of A if,

$$(\lambda I - A)x = 0, x \neq 0$$

- Suppose that $f : R^{m \times n} \to R$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value. Then the **gradient** of f

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

- Suppose that $f : R^n \to R$ is a function that takes a vector in R^n and returns a real number. Then the **Hessian** matrix with respect to x is the $n \times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) \in \mathbb{R}^{n\times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}.$$

Probability Review - Basic Definition

- **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:
  $P(A) \geq 0$ for every A;
  $P(\Omega) = 1$;
  If $A_1, A_2, \cdots$ are disjoint, then $P(U_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
- A **random variable** is a function that maps from the sample space to the reals $(X : \Omega \rightarrow R)$.

Probability Review - Distribution Function

Definition: Suppose X is a random variable, x is a specific value that it can take,
Cumulative distribution function (CDF) is the function $F : R \to [0, 1]$, 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, $f(x) = F'(x)$.

Probability Review - Expectation

Expected Values

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

* To make the above two definitions more explicit: $E[g(X)] = \sum_{X=x \in \mathcal{X}} g(X=x)f(X=x)$, $\mathcal{X}$ is the set of all possible values, e.g. $\{0, 1\}$ when tossing a coin.

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

- $E[a]$ for any constant $a \in \mathbb{R}$.
- $E[af(X)] = aE[f(X)]$ for any constant $a \in R$.
- (Linearity of Expectation) $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.
- $var[X] = E[X^2] - (E[X])^2$.

Example: Mean and Variance of Uniform (n, p)

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x f(x)\mathrm{d}x = \int_a^b x\frac{1}{b-a}\mathrm{d}x = \frac{1}{2(b-a)}\left[x^2\right]_a^b \\
&= \frac{b^2-a^2}{2(b-a)} \\
&= \frac{b+a}{2}
\end{aligned}
$$

$$
\begin{aligned}
V(X) &= E(X^2) - [E(X)]^2 \\
&= \int_a^b x^2.\frac{1}{b-a}\mathrm{d}x - \left(\frac{b+a}{2}\right)^2 = \frac{1}{3(b-a)}\left[x^3\right]_a^b - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{b^3-a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{b^2+ab+a^2}{3} - \frac{b^2+2ab+a^2}{4} \\
&= \frac{(b-a)^2}{12}
\end{aligned}
$$

Probability Review - Common Distribution

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a}$ $\forall x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Definition: **joint cumulative distribution function**

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

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 continous variable.

Conditional Probability and Bayes Rule

Conditional Probability of X given $Y = y$ is

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

Bayes Rule:

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

Chain Rule for multiple random variables:

$$\begin{aligned} f(x_1,x_2,\cdots,x_n) &= f(x_n|x_1,x_2,\cdots,x_{n-1})f(x_1,x_2,\cdots,x_{n-1}) \\ &= f(x_n|x_1,x_2,\cdots,x_{n-1})f(x_{n-1}|x_1,x_2,\cdots,x_{n-2})f(x_1,x_2,\cdots,x_{n-2}) \\ &= \cdots = f(x_1)\prod_{i=2}^{n} f(x_i|x_1,\cdots,x_{i-1}) \end{aligned} \tag{3}$$

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.

Law of Large Numbers

- The **weak law of large numbers** states that the sample average converges in probability towards the expected value

  $$\overline{X}_n \xrightarrow{P} \mu \qquad \text{when } n \to \infty$$

  That is to say that for any positive number ϵ,

  $$\lim_{n\to\infty} \Pr\left(\, |\overline{X}_n - \mu| > \varepsilon \,\right) = 0.$$

- The **strong law of large numbers** states that the sample average converges almost surely to the expected value

  $$\bar{X}_n \xrightarrow{\text{a.s.}} \mu \qquad \text{when } n \to \infty$$

  That is,

  $$\Pr\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

Central Limit Theorem

Central Limit Theorem Suppose $\{X_1, X_2, \cdots\}$ is a sequence of i.i.d. random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{n \to \infty} \mathcal{N}(0, \sigma^2)\right)$$

Where $S_n = \frac{X_1 + \cdots + X_n}{n}$ is the sample average. In probability theory, the central limit theorem (CLT) establishes that, in most situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed.

Optimization - Lagrange multipliers

- In mathematical optimization, the method of Lagrange multipliers is a strategy for finding the local maxima and minima of a function subject to equality constraints.
- Consider an optimization problem:

$$\begin{aligned} \text{minimize} \quad & f(x_1, \cdots, x_n) \\ \text{subject to} \quad & g_k(x_1, \cdots, x_n) = 0, \quad k = 1, \ldots, M \end{aligned} \tag{4}$$

  The Lagrangian takes the form

$$\mathcal{L}(x_1, \cdots, x_n, \lambda_1, \cdots, \lambda_M) = f(x_1, \ldots, x_n) - \sum_{k=1}^{M} \lambda_k g_k(x_1, \ldots, x_n)$$

- Methods of solving optimizaiton using Lagrangian multipliers:

Optimization - Lagrange multipliers

- Solve the following system of equations.

$$
\begin{aligned}
\frac{\partial L(x_1, \cdots, x_n, \lambda_1, \cdots, \lambda_M)}{\partial x_i} &= 0 \text{ , where } i = 1 \cdots n \\
\frac{\partial L(x_1, \cdots, x_n, \lambda_1, \cdots, \lambda_M)}{\partial \lambda_k} &= 0 \text{ , where } k = 1 \cdots M
\end{aligned} \tag{5}
$$

$$g_k(x_1, \cdots, x_n) = 0 \text{ , where } k = 1 \cdots M$$

- Plug in all solutions $x_1, \cdots, x_n$, from the first step into $f(x_1, \cdots, x_n)$ and identify the minimum and maximum values, provided they exist.

Optimization - Lagrange multipliers

Find the extrema of the function $f(x, y) = 2y + x$ subject to the constraint $0 = g(x, y) = y^2 + xy - 1$.
Solution: Set $\mathcal{L}(x, y, \lambda) = f(x, y) + \lambda g(x, y)$, then

$$\begin{aligned} \frac{\partial L}{\partial x} &= 1 + \lambda y \\ \frac{\partial L}{\partial y} &= 2 + 2\lambda y + \lambda x \\ \frac{\partial L}{\partial \lambda} &= y^2 + xy - 1 \end{aligned} \tag{6}$$

Setting these equal to zero, we see from the third equation that $y \neq 0$, and from the first equation that $\lambda = \frac{-1}{y}$, so that from the second equation $0 = \frac{-x}{y}$ implying that $x = 0$. From the third equation, we obtain $y = \pm 1$.

References

- This slide is adapted from course material by Zico Kolter, Chuong Do, Arian Maleki, Tom Do and Ameet Talwalker.

The End