Mathematical Background

Yuh-Jye Lee

Data Science & Machine Intelligence Lab
Dept. of Applied Math @ NCTU

February 21, 2017

Outline

1. Linear Algebra
2. Multi-variable Calculus
3. Probability and Statistics
4. Probability and Inference

Norm

Definition

A norm is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ which must satisfy the following three conditions:

1. $\|x\| \geq 0$, and $\|x\| = 0$ only if $x = 0$,
2. $\|x + y\| \leq \|x\| + \|y\|$,
3. $\|\alpha x\| = |\alpha| \|x\|$.

Variants of Norm

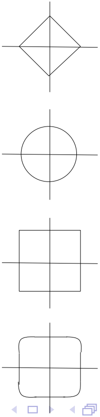
- The most popular vector norms are defined below.
- The closed unit ball $\{x \in \mathbb{R}^n : \|x\| \leq 1\}$ corresponding to each norm is illustrated to the right for the case $n = 2$.

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|,$$

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} = \sqrt{x^\top x},$$

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|,$$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \quad (1 \leq p \leq \infty)$$



Positive Definite Matrices

Definition

An $n \times n$ real symmetric matrix M is positive definite if $z^\top Mz > 0$ for all non-zero vectors $z \in \mathbb{R}^n$.

- Characteristics
  - All eigenvalues λ of M are positive.
  - There exists a unique lower triangular matrix L, with strictly positive diagonal elements, that allows the factorization of M into $M = LL^\top$. This factorization is called Cholesky decomposition.

Eigenvalues and Eigenvectors

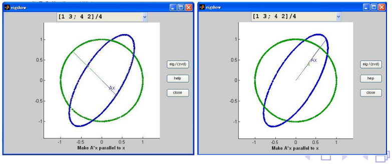
Definition

Given a linear transformation A, a non-zero vector x is defined to be an eigenvector of the transformation if it satisfies the eigenvalue equation

$$Ax = \lambda x$$

for some scalar λ. In this situation, the scalar λ is called an eigenvalue of A corresponding to the eigenvector x.

- You can type *eigshow* in MATLAB to see the graphical demonstration of eigenvalues.



Diagonalization

A matrix $A_{n\times n}$ with n real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and their associated eigenvectors $q_1, q_2, \ldots, q_n$ can be diagonalized as follows:

$$A = Q\Lambda Q^\top,$$

where

$$\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n),\ Q = [q_1|q_2|\ldots q_n]$$

- The eigenvectors are the *principal components*. Extremely important in Machine Learning

Cholesky Factorization

- A matrix decomposition makes $A_{n\times n} = R_{n\times n}^\top R_{n\times n}$, where R is an upper-triangular matrix.
- The matrix A must be positive definite.

$$
\begin{aligned}
A &= \begin{bmatrix} a_{11} & w^t \\ w & K \end{bmatrix} && (1)\\
&= \begin{bmatrix} \alpha & 0 \\ \frac{w}{\alpha} & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & K - \frac{ww^\top}{a_{11}} \end{bmatrix} \begin{bmatrix} \alpha & \frac{w}{\alpha} \\ 0 & I \end{bmatrix} && (2)\\
&= R_1^\top A_1 R_1 && (3)\\
&= R_1^\top R_2^\top \cdots R_m^\top R_m \cdots R_2 R_1 && (4)\\
&= R^\top R && (5)
\end{aligned}
$$

QR Factorization

- A matrix decomposition makes $A_{m\times n} = Q_{m\times m}R_{m\times n}$, where $Q^\top Q = I_{m\times m}$ and R is an upper-triangular matrix.
- QR factorization can be computed by Gram-Schmidt process and Householder transformations.
  - Note: A matrix Q is called *orthogomal matrix* if $Q^\top Q = I$
- For a rectangular matrix:

$$A_{m\times n} = Q_{m\times m}R_{m\times n} = \begin{bmatrix} \hat{Q}_{m\times n} & Q^0_{m\times(m-n)} \end{bmatrix} \begin{bmatrix} \hat{R}_{n\times n} \\ 0 \end{bmatrix} = \hat{Q}_{m\times n}\hat{R}_{n\times n}$$

- $A = \hat{Q}\hat{R}$ is the reduced QR factorization

Singular Value Decomposition (SVD)

- A matrix decomposition makes $A_{m\times n} = U_{m\times m}\Sigma_{m\times n}V_{n\times n}^\top$, where $U^\top U = I_{m\times m}$ and $V^\top V = I_{n\times n}$.
- U and V are the eigenvectors of $AA^\top$ and $A^\top A$ respectively.
- For a rectangular matrix:

$$A = U\Sigma V^\top = \begin{bmatrix} \hat{U}_{m\times n} & U^0_{m\times(m-n)} \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_{n\times n} \\ 0 \end{bmatrix} V = \hat{U}_{m\times n}\hat{\Sigma}_{n\times n}V_{n\times n}^\top$$

- $A = \hat{U}\hat{\Sigma}V^\top$ is the reduced SVD.
- SVD is the Latent Semantic Indexing (LSI) in Text Mining when A is a *term by document* matrix

Least Squares Problem

- Given $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$, a linear system with $m > n$:

$$Aw = y, \tag{6}$$

  is called an overdetermined linear system.

- In general, an overdetermined linear system has no solution. An approximated solution can be obtained by solving the following minimization problem.

$$\min_{w \in \mathbb{R}^n} r^\top r = \min_{w \in \mathbb{R}^n} \|r\|_2^2 = \min_{w \in \mathbb{R}^n} \sum_{i=1}^{m} (y_i - A_i w)^2, \tag{7}$$

  where $r = y - Aw \in \mathbb{R}^m$ is the *residual*.

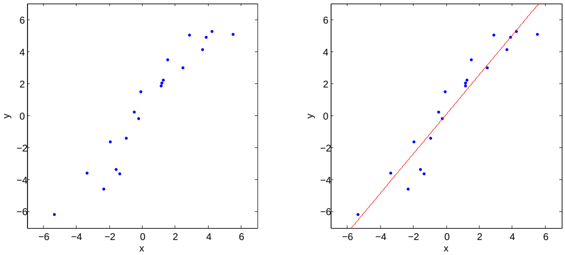
- The minimization problem (7) is the formulation of *least squares problem*.

Example: Data Fitting

Suppose we want to fit the data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$$

with a straight line $y = w_0 + w_1 x$.



Example: Data Fitting

This problem can be expressed as the following overdetermined linear system:

$$
\begin{aligned}
y_1 &= w_0 + w_1 x_1 \\
y_2 &= w_0 + w_1 x_2 \\
&\vdots \\
y_m &= w_0 + w_1 x_m,
\end{aligned}
$$

or

$$
\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \tag{8}
$$

or

$$
Aw = y.
$$

A vector w minimizes the residual norm $\|r\|_2 = \|y - Aw\|_2$, thereby solving the least squares problem if and only if $r \perp \text{range}(A)$, that is,

$$A^\top r = 0$$

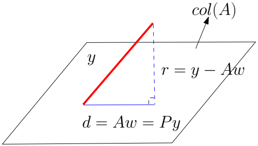
or equivalently,

$$A^\top A w = A^\top y,$$

or equivalently,

$$Py = Aw,$$

where $P = A(A^\top A)^{-1} A^\top$ is a orthogonal projection and w is unique iff A is full rank ($w = (A^\top A)^{-1} A^\top y$).



Outline

1. Linear Algebra
2. Multi-variable Calculus
3. Probability and Statistics
4. Probability and Inference

Gradient

Definition

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable function. The gradient of function f at a point $x \in \mathbb{R}^n$ is defined as

$$\nabla f(x) = [\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n}] \in \mathbb{R}^n$$

- The gradient vector $\nabla f(x)$ gives the direction of fastest increase of f.
- Example

$$\begin{aligned} f(x_1, x_2) &= x_1^2 + x_2^2 - 2x_1 + 4x_2 \\ \nabla f(x_1, x_2) &= \begin{bmatrix} 2x_1 - 2 & 2x_2 + 4 \end{bmatrix} \end{aligned}$$

Hessian

Definition

If $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function. The Hessian matrix of f at a point $x \in \mathbb{R}^n$ is defined as

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

- Hessian matrix describes the local curvature of a function
- Example

$$\begin{aligned} f(x_1, x_2) &= x_1^2 + x_2^2 - 2x_1 + 4x_2 \\ \nabla^2 f(x) &= \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \end{aligned}$$

Connection to Maximum and Minimum Values

First-Order Necessary Conditions

If x^* is a local minimizer and f is continuously differentiable in an open neighborhood of x^*, then $\nabla f(x^*) = 0$.

Second-Order Necessary Conditions

If x^* is a local minimizer and $\nabla^2 f$ exists and is continuous in an open neighborhood of x^*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.

Second-Order Sufficient Conditions

Suppose that $\nabla^2 f$ is continuous in an open neighborhood of x^* and that $\nabla f(x^*) = 0$, and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimizer of f.

Revisit Least Squares Problem

- Given $A \in \mathbb{R}^{m\times n}$ and $y \in \mathbb{R}^m$, a linear system with $m > n$:

$$Aw = y, \tag{9}$$

  is called an overdetermined linear system.

- Try to find an approximation solution with the "smallest residual"

$$\min_{w\in\mathbb{R}^n} \|r\|_2^2 = \min_{w\in\mathbb{R}^n} \sum_{i=1}^{m} (y_i - A_i w)^2 = \min_{w\in\mathbb{R}^n} f(w). \tag{10}$$

- Let $\nabla f(w) = \mathbf{0}$ we can have the *normal equation*

Outline

Random Variable

Definition

A *random variable* is a real-valued function for which domain is a sample space

- Example

  For a coin toss, the possible outcome is head or tail. The number of heads appearing in one fair coin toss can be described using the following random variable:

$$X = \begin{cases} 1, & \text{if head} \\ 0, & \text{if tail} \end{cases}$$

  with probability function given by:

$$P(X = x) = \begin{cases} \frac{1}{2}, & \text{if } x = 1 \\ \frac{1}{2}, & \text{if } x = 0 \\ 0, & \text{sotherwise} \end{cases}$$

Probability Distribution

Definition

If X is discrete random variable, the function given by $P(X = x)$ for each x within the range of X is called probability distribution of X.

- Example
  Let the random variable X be denoted as the total number of heads. The probability distribution of heads obtained in the four tosses of a fair coin can be written as follows:

$$P(X = x) = \frac{\binom{4}{x}}{2^4}, \text{ for } x = 0, 1, 2, 3, 4.$$

Probability Density Distribution

Definition

A function with values $f(x)$, defined over the set of all real numbers, is called a probability density function of the continuous random variable X if and only if

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

for any real constants a and b with $a \leq b$

- Example
  The p.d.f of normal distribution is defined as follows:
  $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2},$$
  where μ is the mean and σ is the standard deviation.

Conditional Probability

Definition

The conditional probability of an event A, given that an event B has occurred, is equal to

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

- Example
  Suppose that a fair die is tossed once. Find the probability of a 1 (event A), given an odd number was obtained (event B).

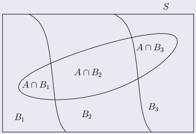
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}$$

- Restrict the sample space on the event B

Theorem

Assume that $\{B_1, B_2, \ldots, B_k\}$ is a partition of S such that $P(B_i) > 0$, for $i = 1, 2, \ldots, k$. Then

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i).$$



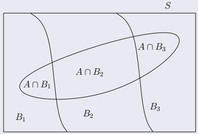
- Note that $\{B_1, B_2, \ldots, B_k\}$ is a partition of S if
  1. $S = B_1 \cup B_2 \cup \ldots \cup B_k$
  2. $B_i \cap B_j = \emptyset$ for $i \neq j$

Bayes' Rule

Bayes' Rule

Assume that $\{B_1, B_2, \ldots, B_k\}$ is a partition of S such that $P(B_i) > 0$, for $i = 1, 2, \ldots, k$. Then

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}.$$



Expected Value

Definition

If X is a discrete random variable and $P(X = x)$ is the value of its probability distribution at x, the expected value of X is

$$\mu = E(X) = \sum_{x} x \cdot P(X = x).$$

Correspondingly, if X is a continuous random variable and $f(x)$ is the value of its probability density at x, the expected value of X is

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx.$$

- $E(aX + bY) = aE(X) + bE(Y)$, *linear operator*

Variance
Measures of how far a set of numbers are spread out

Definition

If X is a discrete random variable and $P(X = x)$ is the value of its probability distribution at x, the expected value of X is

$$Var(X) = E([X - E(X)]^2) = \sum_x (x - \mu)^2 \cdot P(X = x).$$

Correspondingly, if X is a continuous random variable and $f(x)$ is the value of its probability density at x, the expected value of X is

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx.$$

- $Var(X) = E(X^2) - (E(X))^2$

Bernoulli Distribution

A trial is performed whose outcome is either a "success" or a "failure". The random variable X is a 0/1 indicator variable and takes the value 1 for a success outcome and is 0 otherwise. p is the probability that the result of trail is a success. Then

$$P(X = 1) = p \text{ and } P(X = 0) = 1 - p$$

which can equivalently be written as

$$P(X = i) = p^i (1 - p)^{1-i}, \ i = 0, 1$$

Tossing a *fair* coin, the parameter $p = 0.5$. If X is Bernoulli,

1. $E(X) = p$,
2. $Var(X) = p(1 - p)$
3. Who knows p?

Probability and Inference

- The outcome of tossing a coin is $\{Heads, Tails\}$
- We use a random variable $X \in \{0,\ 1\}$ to indicate the outcome
- Suppose that we have a random sample: $\mathbf{X} = \{x^t\}_{t=1}^N$
- How to *estimate* the parameter p?

Maximum Likelihood Estimation

Likelihood Function

The probability to *observe* the random sample $\mathbf{X} = \{x^t\}_{t=1}^N$ is

$$\prod_{t=1}^{N} p^{x^t}(1-p)^{1-x^t}$$

Why don't we choose the parameter p which will maximize the probability for observing the random sample $\mathbf{X} = \{x^t\}_{t=1}^N$?

Based on MLE, we will choose the parameter p

$$p = \frac{\sum_{t=1}^{N} x^t}{N}$$

Sample Mean, Variance, and Standard deviation

Sample Mean

The mean of a sample of n measured responses $y_1, y_2, \ldots, y_n$ is given by

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

The corresponding population mean is denoted by μ.

Sample Variance

The variance of a sample of measurements $y_1, y_2, \ldots, y_n$ is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

The corresponding population variance is denoted by σ^2.

Applying Baye's Rule to Classification

Credit Cards Scoring: Low-risk vs. High-risk

- According to the past transactions, some customers are low-risk in that they paid back their loan and the bank profited from them and other customers are high-risk in that they defaulted.
- We would like to *learn* the class "*high-risk customer*"
- We observe customer's *yearly income* and *savings*, which we represent by two *random variables* X_1 and X_2
- The *credibility of a customer* is denoted by a *Bernoulli* random variable C where $C = 1$ indicates a high-risk customer and $C = 0$ indicated a low-risk customer

Applying Baye's Rule to Classification

How to make the decision when a new application arrives?

- When a new application arrives with $X_1 = x_1$ and $X_2 = x_2$
- If we know the probability of C *conditioned on* the observation $X = [x_1,\ x_2]$ our decision will be
    - $C = 1$ if $P(C = 1|[x_1,\ x_2]) > 0.5$
    - $C = 0$ otherwise
- The probability of error we made based on this rule is

$$1 - \max\{P(C = 1|[x_1,\ x_2]), P(C = 0|[x_1,\ x_2])\} < 0.5$$

- Please note $P(C = 1|[x_1,\ x_2]) + P(C = 0|[x_1,\ x_2]) = 1$

The *Posterior Probability*: $P(C|\mathbf{x}) = \frac{P(C)P(\mathbf{x}|C)}{P(\mathbf{x})}$

- $P(C = 1)$ is called the *prior probability* that $C = 1$
- In our example, it corresponds to a probability that a customer is high-risk, *regardless* of the $\mathbf{x}$ value.
- It is called the *prior probability* because it is the knowledge we have *before* looking at the observation $\mathbf{x}$
- $P(\mathbf{x}|C)$ is called the *class likelihood* and is the *conditional probability* that an *event belonging to the class C* has the associated observation value $\mathbf{x}$
- $P(\mathbf{x})$, the *evidence* is the probability that an observation $\mathbf{x}$ to be seen, regardless of whether it is a positive or negative example

All above information can be extracted from the past transactions *(historical data)*

The *Posterior Probability*: $P(C|\mathbf{x}) = \frac{P(C)P(\mathbf{x}|C)}{P(\mathbf{x})}$

- Because of normalization by the evidence, the posteriors sum up to 1
- In our example, $P(X_1, X_2)$ is called the *joined probability* of two random variables X_1 and X_2
- Under the assumption, these two random variables X_1 and X_2 are *probability independent*, we have $P(X_1, X_2) = P(X_1)P(X_2)$
- It is one of key assumptions of *Naive Bayes' Classifier*
- Although it is *over simplified* the problem it is very easy to use for real applications

Extend to Multi-class classification

- We have K mutually and exhaustive classes; $C_i, \; i = 1, 2, \ldots, K$
- For example, in *optical digit recognition*, the input is a *bitmap image* and there are 10 classes
- We can think of that these K classes define a *partition* of the *input space*
- Please refer to the slides of the *Partition Theorem* and *Baye's Rule*
- The Bayes' classifier choose the class with the highest posterior probability; that is we choose C_i if

$$P(C_i|\mathbf{x}) = \max_k P(C_k|\mathbf{x})$$

- Question: Is it very important to have $P(\mathbf{x})$, the evidence?