CS 6316 Machine Learning

Review of Linear Algebra and Probability

Yangfeng Ji

Department of Computer Science
University of Virginia

Overview

1. Course Information
2. Basic Linear Algebra
3. Probability Theory
4. Statistical Estimation

Course Information

Instructors

- Yangfeng Ji
  - Office hour: Wednesday 11 AM - 12 PM
  - Office: Rice 510
- Hanjie Chen (TA)
  - Office hour: Tuesday and Thursday 1 PM – 2 PM
  - Office: Rice 442
- Kai Lin (TA)
  - Office hour: TBD

Goal

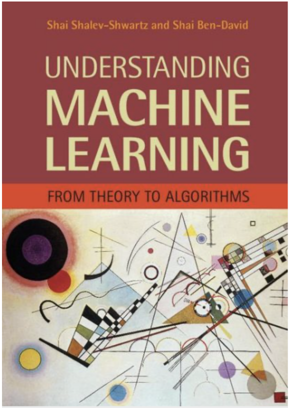
Understand the basic concepts and models from the computational perspective

To

- provide a wide coverage of basic topics in machine learning
  - Example: PAC learning, linear predictors, SVM, boosting, kNN, decision trees, neural networks, etc
- discuss a few fundamental concepts in each topic
  - Example: learnability, generalization, overfitting/underfitting, VC dimension, max margins methods, etc.

Textbook

Shalev-Shwartz and Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. 2014[1]



[1] https://www.cse.huji.ac.il/~shais/UnderstandingMachineLearning/index.html

Outline

This course will cover the basic materials on the following topics

1. Learning theory
2. Linear classification and regression
3. Model selection and validation
4. Boosting and support vector machines
5. Neural networks
6. Clustering and dimensionality reduction

Outline (II)

The following topics will not be the emphasis of this course

- Statistical modeling
  - Statistical Learning and Graphical Models by Farzad Hassanzadeh
- Deep learning
  - Deep Learning for Visual Recognition by Vicente Ordonez-Roman

Reference Courses

For fans of machine learning:

- Shalev-Shwartz. Understanding Machine Learning. 2014
- Mohri. Foundations of Machine Learning. Fall 2018

Reference Books

For fans of machine learning:

- Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning (2nd Edition). 2009
- Murphy. Machine Learning: A Probabilistic Perspective. 2012
- Bishop. Pattern Recognition and Machine Learning. 2006
- Mohri, Rostamizadeh, and Talwalkar. Foundations of Machine Learning. 2nd Edition. 2018

Homework and Grading Policy

- Homeworks (75%)
  - Five homeworks, each of them worth 15%
- Final project (22%)
  - Project proposal: 5%
  - Midterm report: 5%
  - Final project presentation: 6%
  - Final project report: 6%
- Class attendance (3%): we will take attendance at three randomly-selected lectures. Each is worth 1%

Grading Policy

The final grade is threshold-based instead of percentage-based

Point range	Letter grade
[99 100]	A+
[94 99)	A
[90 94)	A-
[88 90)	B+
[83 88)	B
[80 83)	B-
[74 80)	C+
[67 74)	C
[60 67)	C-

Late Penalty

- Homework submission will be accepted up to 72 hours late, with 20% deduction per 24 hours on the points as a penalty
- It is usually better if students just turn in what they have in time
- Submission will not be accepted if more than 72 hours late
- Do not submit the wrong homework — late penalty will be applied if resubmit after deadline

Violation of the Honor Code

Plagiarism, examples are

- in a homework submission, copying answers from others directly (even, with some minor changes)
- in a report, copying texts from a published paper (even, with some minor changes)
- in a code, using someone else's functions/implementations without acknowledging the contribution

Webpages

- Course webpage

  `http://yangfengji.net/uva-ml-course/`

  which contains all the information you need about this course.
- Piazza

  `https://piazza.com/virginia/spring2020/cs6316/home`

Basic Linear Algebra

Consider the following system of equations

$$\begin{aligned} 4x_1 - 5x_2 &= -13 \\ -2x_1 + 3x_2 &= 9 \end{aligned} \tag{1}$$

In matrix notation, it can be written as a more compact from

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b} \tag{2}$$

with

$$\mathbf{A} = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} -13 \\ 9 \end{bmatrix} \tag{3}$$

Basic Notations

$$\mathbf{A} = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$$

- $\mathbf{A} \in \mathbb{R}^{m \times n}$: a matrix with m rows and n columns
  - The element on the i-th row and the j-th column is denoted as $a_{i,j}$
- $\boldsymbol{x} \in \mathbb{R}^n$: a vector with n entries. By convention, an n-dimensional vector is often thought of as matrix with n rows and 1 column, known as a column vector.
  - The i-th element is denoted as x_i

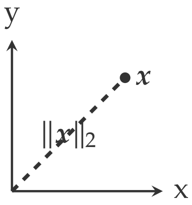
Vector Norms

- A norm of a vector $\|\boldsymbol{x}\|$ is informally a measure of the "length" of the vector.
- Formally, a norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies four properties
  1. $f(\boldsymbol{x}) \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^n$
  2. $f(\boldsymbol{x}) = 0$ if and only if $\boldsymbol{x} = 0$
  3. $f(a\boldsymbol{x}) = |a| \cdot f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^n$
  4. $f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

ℓ_2 Norm

The ℓ_2 norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{4}$$



Exercise: prove ℓ_2 norm satisfies all four properties

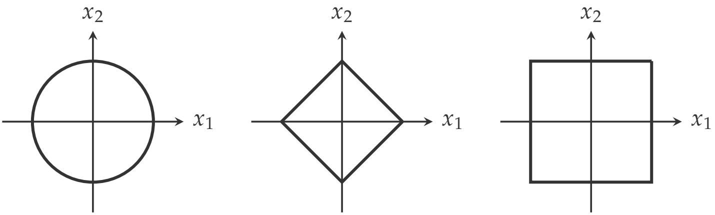
ℓ_1 Norms

The ℓ_1 norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| \tag{5}$$

Quiz

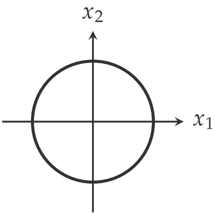
For a two-dimensional vector $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$, which of the following plot is $\|\boldsymbol{x}\|_1 = 1$?



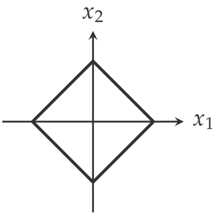
(a) (b) (c)

Quiz

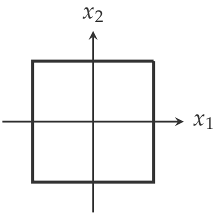
For a two-dimensional vector $\boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2$, which of the following plot is $\|\boldsymbol{x}\|_1 = 1$? Answer: (b)



(d)



(e)



(f)

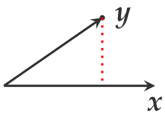
Dot Product

The dot product of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is defined as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i \tag{6}$$

where $\boldsymbol{x}^\top$ is the transpose of $\boldsymbol{x}$.

- $\|\boldsymbol{x}\|_2^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle$
- If $\boldsymbol{x} = (0, 0, \ldots, \underbrace{1}_{x_i}, \ldots, 0)$, then $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = y_i$
- If $\boldsymbol{x}$ is an unit vector ($\|\boldsymbol{x}\|_2 = 1$), then $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is the projection of $\boldsymbol{y}$ on the direction of $\boldsymbol{x}$



Cauchy-Schwarz Inequality

For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \tag{7}$$

with equality if and only if $\boldsymbol{x} = \alpha \boldsymbol{y}$ with $\alpha \in \mathbb{R}$

Proof:

Let $\tilde{\boldsymbol{x}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$ and $\tilde{\boldsymbol{y}} = \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2}$, then $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$ are both unit vectors.

Based on the geometric interpretation on the previous slide, we have

$$\langle \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}} \rangle \leq 1 \tag{8}$$

if and only if $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{y}}$.

Frobenius Norm

The Forbenius norm of a matrix $\mathbf{A} = [a_{i,j}] \in \mathbb{R}^{m \times n}$ denoted by $\| \cdot \|_F$ is defined as

$$\|\mathbf{A}\|_F = \Big(\sum_i \sum_j a_{i,j}^2 \Big)^{1/2} \tag{9}$$

- The Frobenius norm can be interpreted as the ℓ_2 norm of a vector when treating $\mathbf{A}$ as a vector of size mn.

Two Special Matrices

- The identity matrix, denoted as $\mathbf{I} \in \mathbb{R}^{n\times n}$], is a square matrix with ones on the diagonal and zeros everywhere else.

$$\mathbf{I} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \tag{10}$$

- A diagonal matrix, denoted as $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, is a matrix where all non-diagonal elements are 0.

$$\mathbf{D} = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix} \tag{11}$$

Inverse

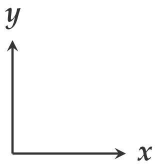
The *inverse* of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is denoted as $\mathbf{A}^{-1}$, which is the unique matrix such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1} \tag{12}$$

- Non-square matrices do not have inverses (by definition)
- Not all square matrices are invertible
- The solution of the linear equations in Eq. (1) is $\boldsymbol{x} = \mathbf{A}^{-1}\boldsymbol{b}$

Orthogonal Matrices

- Two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ are orthogonal if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$



- A square matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ is orthogonal, if all its columns are orthogonal to each other *and* normalized (orthonormal)

$$\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = 0, \|\boldsymbol{u}_i\| = 1, \|\boldsymbol{u}_j\| = 1 \tag{13}$$

  for $i, j \in [n]$ and $i \neq j$
- Furthermore, $\mathbf{U}^\top \mathbf{U} = \mathbf{I} = \mathbf{U}\mathbf{U}^\top$, which further implies $\mathbf{U}^{-1} = \mathbf{U}^\top$

Symmetric Matrices

A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is defined as

$$\mathbf{A}^\top = \mathbf{A} \tag{14}$$

or, in other words,

$$a_{i,j} = a_{j,i} \quad \forall i, j \in [n] \tag{15}$$

Comments

- The identity matrix $\mathbf{I}$ is symmetric
- A diagonal matrix is symmetric

Eigen Decomposition

Every symmetric matrix $\mathbf{A}$ can be decomposed as

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\top} \tag{16}$$

with

- $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$ as a diagonal matrix (Slide 25)
- $\mathbf{Q}$ is an orthogonal matrix (Slide 27)
- *Exercise*: if $\mathbf{A}$ is invertible, show $\mathbf{A}^{-1} = \mathbf{U}\Lambda^{-1}\mathbf{U}^{\top}$ with $\Lambda^{-1} = \mathrm{diag}(\frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_n})$

Symmetric Positive Semidefinite Matrices

A symmetric matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ is positive semidefinite if and only if

$$\boldsymbol{x}^\top \mathbf{P} \boldsymbol{x} \geq 0 \tag{17}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$.

Symmetric Positive Semidefinite Matrices

A symmetric matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ is positive semidefinite if and only if

$$\boldsymbol{x}^\top \mathbf{P} \boldsymbol{x} \geq 0 \tag{17}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$.

Eigen decomposition (Slide 29) of $\mathbf{P}$ as

$$\mathbf{P} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^\top \tag{18}$$

with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and

$$\lambda_i \geq 0 \tag{19}$$

Symmetric Positive Definite Matrices

A symmetric matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ is positive definite if and only if

$$\boldsymbol{x}^\top \mathbf{P} \boldsymbol{x} > 0 \tag{20}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$.

- Eigen values of $\mathbf{P}$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ with

$$\lambda_i > 0 \tag{21}$$

- *Exercise*: if one of the eigen values $\lambda_i < 0$, show that you can also find a vector $\boldsymbol{x}$ such that $\boldsymbol{x}^\top \mathbf{P} \boldsymbol{x} < 0$

Quiz

The identity matrix **I** is

- a diagonal matrix?
- a symmetric matrix?
- an orthogonal matrix?
- a positive (semi-)definite matrix?

Further reference [Kolter and Do, 2015]

Quiz

The identity matrix $\mathbf{I}$ is

- a diagonal matrix? ✓
- a symmetric matrix? ✓
- an orthogonal matrix? ✓
- a positive (semi-)definite matrix? ✓

Further reference [Kolter and Do, 2015]

Probability Theory

What is Probability?

The probability of landing heads is 0.52

Two interpretations

Frequentist Probability represents the *long-run frequency* of an event

- If we flip the coin many times, we expect it to land heads about 52% times

Two interpretations

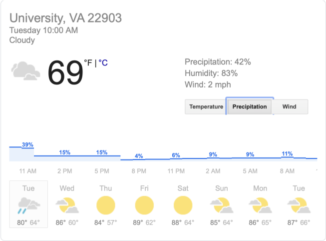
Frequentist Probability represents the *long-run frequency* of an event

- If we flip the coin many times, we expect it to land heads about 52% times

Bayesian Probability quantifies our *(un)certainty* about an event

- We believe the coin is 52% of chance to land head on the next toss

Example scenarios of Bayesian interpretation of probability:



Binary Random Variables

- **Event** X. Such as
  - *the coin will lead head on the next toss*
  - *it will rain tomorrow*
- Sample space of $X \in \{\text{false}, \text{true}\}$ or for simplicity $\{0, 1\}$

Binary Random Variables

- **Event** X. Such as
  - *the coin will lead head on the next toss*
  - *it will rain tomorrow*
- Sample space of $X \in \{\text{false}, \text{true}\}$ or for simplicity $\{0, 1\}$
- Probability $P(X = x)$ or $P(x)$
- Let X be the event that *the coin will lead head on the next toss*, then the probability from the previous example is

$$P(X = 1) = 0.52 \tag{22}$$

Bernoulli Distribution

Given the binary random variable X and its sample space as $\{0, 1\}$

$$P(X = x) = \theta^x (1 - \theta)^{1-x}$$

with a single parameter θ as

$$\theta = P(X = 1)$$

Jacob Bernoulli

Tossing a Coin Twice?

- Let X be the number of heads
- Sample space of $X \in \{0, 1, 2\}$

Tossing a Coin Twice?

- Let X be the number of heads
- Sample space of $X \in \{0, 1, 2\}$
- Assume we use the same coin, the probability distribution of X
  - $P(X = 0) = (1 - \theta)^2$

Tossing a Coin Twice?

- Let X be the number of heads
- Sample space of $X \in \{0, 1, 2\}$
- Assume we use the same coin, the probability distribution of X
  - $P(X = 0) = (1 - \theta)^2$
  - $P(X = 2) = \theta^2$

Tossing a Coin Twice?

- Let X be the number of heads
- Sample space of $X \in \{0, 1, 2\}$
- Assume we use the same coin, the probability distribution of X
  - $P(X = 0) = (1 - \theta)^2$
  - $P(X = 2) = \theta^2$
  - $P(X = 1) = \theta(1 - \theta) + (1 - \theta)\theta = 2\theta(1 - \theta)$

General Case: Binomial Distribution

Consider a general case, in which we toss the coin n times, then the random variable Y can be formulated as a binomial distribution

$$P(Y = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{23}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

is the binomial coefficient and

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 1$$

Tossing a Dice

How to define the corresponding random variable?

- $X \in \{1, 2, 3, 4, 5, 6\}$
- $\boldsymbol{X} \in \{100000, 010000, 001000, 000100, 000010, 000001\}$

Categorical Distribution

$$P(\boldsymbol{X} = \boldsymbol{x}) = \prod_{k=1}^{6} (\theta_k)^{x_k} \tag{24}$$

where

- $\boldsymbol{x} = (x_1, x_2, \ldots, x_6)$
- $x_k \in \{0, 1\}$, and
- $\{\theta_k\}_{k=1}^{6}$ are the parameters of this distribution, which is also the probability of side k showing up.

Multinomial Distribution

Repeat the previous event n times, the corresponding probability distribution is modeled as

$$P(\boldsymbol{X} = \boldsymbol{x}) = \binom{n}{x_1 \cdots x_K} \prod_{k=1}^{K} \theta_k^{x_k} \tag{25}$$

where $\boldsymbol{x} = (x_1, \ldots, x_K)$ and each $x_k \in \{0, 1, 2, \ldots, n\}$ indicates the number of times that side k showing up.

$$\binom{n}{x_1 \cdots x_K} = \frac{n!}{x_1! \cdots x_K!}$$

The sum of $\{x_k\}$ follows the constraint:

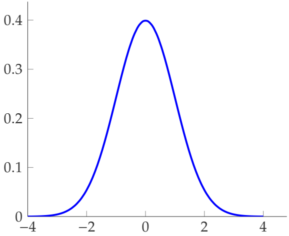
$$\sum_{k=1}^{K} x_k = n$$

Gaussian Distribution

A random variable $X \in \mathbb{R}$ is said to follow a normal (or Gaussian) distribution $\mathcal{N}(\mu, \sigma^2)$ if its probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{26}$$

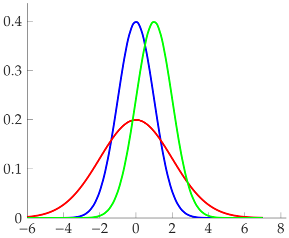
- μ: mean
- σ^2: variance
- Probability of $X \in [a, b]$: $P(a \leq X \leq b) = \int_a^b f(x)dx$



Gaussian Distribution (II)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{27}$$

There examples of Gaussian distributions



- Blue: $\mathcal{N}(0,1)$ (standard normal distribution)
- Red: $\mathcal{N}(0,2)$
- Green: $\mathcal{N}(1,1)$

Probability of Two Random Variables

Modeling two random variables together with a **joint** distribution

$$P(X \cap Y) \tag{28}$$

or, sometimes for simplicity

$$P(X, Y) \tag{29}$$

Related concepts

- Independence
- Conditional probability and chain rule
- Bayes rule

Independence

Definition Two random variable X and Y are independent with each other, if we can represent the joint probability as the product of their marginal distributions for *any* values of X and Y, or mathematically,

$$P(X \cap Y) = P(X) \cdot P(Y) \tag{30}$$

Marginal distributions

$$P(X) = \sum_{Y} P(X \cap Y) \tag{31}$$

$$P(Y) = \sum_{X} P(X \cap Y) \tag{32}$$

Independence

Definition Two random variable X and Y are independent with each other, if we can represent the joint probability as the product of their marginal distributions for *any* values of X and Y, or mathematically,

$$P(X \cap Y) = P(X) \cdot P(Y) \tag{30}$$

Marginal distributions

$$P(X) = \sum_{Y} P(X \cap Y) \tag{31}$$

$$P(Y) = \sum_{X} P(X \cap Y) \tag{32}$$

- X: whether it is cloudy
- Y: whether it will rain

$P(X \cap Y)$	$X = 0$	$X = 1$
$Y = 0$	0.35	0.15
$Y = 1$	0.05	0.45

Conditional Probability

Conditional probability of Y given X

$$P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} \tag{33}$$

Example: document classification

- X: a document
- Y: the label of this document

A special case: if X and Y are independent

$$P(Y \mid X) = P(Y) \tag{34}$$

Intuitively, it means *Knowing X does not provide any new information about Y*

Conditional Probability

- X: whether it is cloudy
- Y: whether it will rain

$P(X \cap Y)$	$X = 0$	$X = 1$
$Y = 0$	0.35	0.15
$Y = 1$	0.05	0.45

- $P(Y \mid X = 1)$:
  - $P(Y = 0 \mid X = 1) = 0.25$,
  - $P(Y = 1 \mid X = 1) = 0.75$

Conditional Probability

$P(X \cap Y)$	$X = 0$	$X = 1$
$Y = 0$	0.35	0.15
$Y = 1$	0.05	0.45

- X: whether it is cloudy
- Y: whether it will rain

- $P(Y \mid X = 1)$:
  - $P(Y = 0 \mid X = 1) = 0.25$,
  - $P(Y = 1 \mid X = 1) = 0.75$
- $P(Y)$: $P(Y = 0) = P(Y = 1) = 0.5$

Multivariate Gaussian

The probability density function of a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined as

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right) \tag{35}$$

where

- $\boldsymbol{\mu}$ is the n-dimensional mean vector and
- $\boldsymbol{\Sigma}$ is the $n \times n$ covariance matrix.

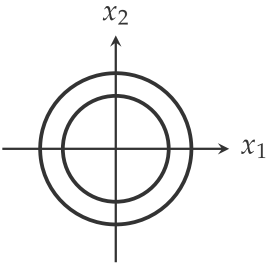
Covariance Matrix Σ

Assume $\boldsymbol{\mu} = 0$, the probability density function is

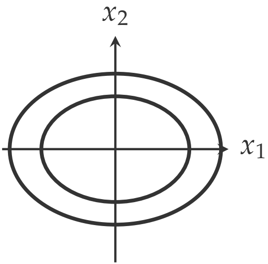
$$f(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right) \tag{36}$$

In general, $\boldsymbol{\Sigma}$ is required to be a symmetric positive definite matrix

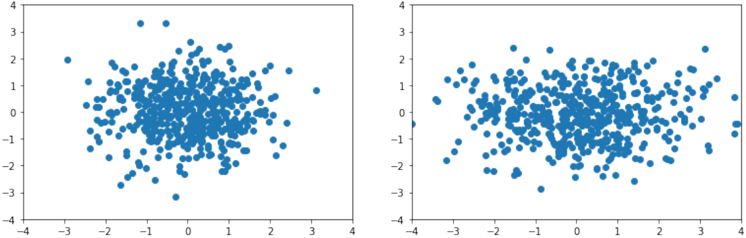
$\boldsymbol{\Sigma} = \mathbf{I}$



$\boldsymbol{\Sigma} = \mathrm{diag}(2, 1)$



Sampling from Gaussians



(a) (b)

(a) : $\boldsymbol{\Sigma} = \mathbf{I}$

(b) : $\boldsymbol{\Sigma} = \mathrm{diag}(2, 1)$

Exercise: Sample from an arbitrary Gaussian distribution

Sum Rule

Given two random variables X and Y, without any additional assumption we have

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y) \qquad (37)$$

- If X and Y are independent, then

$$P(X \cap Y) = 0 \quad \text{and} \quad P(X \cup Y) = P(X) + P(Y) \qquad (38)$$

- *Exercise*: Prove the following inequality by generalizing the sum rule in

$$P(\cup_{i=1}^{n} X_i) \leq \sum_{i=1}^{n} P(X_i) \qquad (39)$$

This inequality is called the union bound.

Chain Rule

Any joint probability of two random variable can be decomposed as

$$P(X, Y) = P(X) \cdot P(Y \mid X) = P(Y) \cdot P(X \mid Y) \tag{40}$$

No independence assumption is needed

Chain Rule

Any joint probability of two random variable can be decomposed as

$$P(X, Y) = P(X) \cdot P(Y \mid X) = P(Y) \cdot P(X \mid Y) \tag{40}$$

No independence assumption is needed

The chain rule can be easily generalized

$$\begin{aligned} P(X_1, X_2, \cdots, X_k) &= P(X_1)P(X_2, \cdots, X_k \mid X_1) \\ &= P(X_1)P(X_2 \mid X_1)P(X_3, \cdots, X_k \mid X_2, X_1) \\ &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2, X_1) \cdots \\ &\quad P(X_k \mid X_1, \cdots, X_{k-1}) \end{aligned} \tag{41}$$

Inverse Probability

Given

- $P(Y)$: prior probability, and
- $P(X \mid Y)$: conditional probability of X given Y,

we can compute the probability $P(Y \mid X)$ using Bayes' rule as

$$P(Y \mid X) = \frac{P(Y)P(X \mid Y)}{P(X)} \tag{42}$$

where

$$P(X) = \sum_{Y} P(Y)P(X \mid Y) \tag{43}$$

Example: The burglar alarm

Two random variables, alarm A and burglar B

- $P(A = 1 \mid B = 1) = 0.99$: burglar happens, alarm rings
- $P(A = 1 \mid B = 0) = 0.001$: burglar does not happen, alarm rings
- $P(B = 1) = 0.01$: burglar rate

Question: if the alarm rang, what is the probability of a burglar happened?

$$P(B = 1 \mid A = 1) \tag{44}$$

Example: The burglar alarm (II)

- $P(A = 1 \mid B = 1) = 0.99$: burglar happens $\Rightarrow$ alarm rings
- $P(A = 1 \mid B = 0) = 0.001$: burglar does not happen $\Rightarrow$ alarm rings
- $P(B = 1) = 0.01$: burglar rate

Question: if the alarm rang, what is the probability of a burglar happened?

$$
\begin{aligned}
&P(B = 1 \mid A = 1) \\
&\qquad = \frac{P(B = 1)P(A = 1 \mid B = 1)}{P(A = 1 \mid B = 1)P(B = 1) + P(A = 1 \mid B = 0)P(B = 0)} \\
&\qquad = \frac{0.01 \times 0.99}{(0.01 \times 0.99) + (0.001 \times (1 - 0.01))} \\
&\qquad \approx 0.91
\end{aligned}
$$

Example: The burglar alarm (II)

- $P(A = 1 \mid B = 1) = 0.99$: burglar happens $\Rightarrow$ alarm rings
- $P(A = 1 \mid B = 0) = 0.001$: burglar does not happen $\Rightarrow$ alarm rings
- $P(B = 1) = 0.01$: burglar rate

Question: if the alarm rang, what is the probability of a burglar happened?

$$
\begin{aligned}
P(B = 1 \mid A = 1) & \\
&= \frac{P(B = 1)P(A = 1 \mid B = 1)}{P(A = 1 \mid B = 1)P(B = 1) + P(A = 1 \mid B = 0)P(B = 0)} \\
&= \frac{0.01 \times 0.99}{(0.01 \times 0.99) + (0.001 \times (1 - 0.01))} \\
&\approx 0.91
\end{aligned}
$$

Further Question: What if $P(A = 1 \mid B = 0) = 0.01$?

Expectation

The expectation or expected value of a function $h(x)$ with respect to a probability distribution $P(X)$ is defined as

$$E\left[h(x)\right] = \sum_{x} P(x)h(x) \tag{45}$$

Expectation

The expectation or expected value of a function $h(x)$ with respect to a probability distribution $P(X)$ is defined as

$$E[h(x)] = \sum_{x} P(x)h(x) \tag{45}$$

The number of ice creams [Eisenstein, 2018]

- If it is sunny, Lucia will eat four ice creams
- If it is rainy, she will eat only one ice cream
- There is a 90% chance it will be rainy

The expected number of ice creams she will eat is

$$(1 - 0.9) \times 4 + 0.9 \times 1 = 1.3 \tag{46}$$

Mean

- Let $h(x) = x$, the expectation is the mean value of the random variable X (discrete random variable)

$$E[X] = \sum_x xP(x) \tag{47}$$

or, (continuous random variable)

$$E[X] = \int_x x f(x) \tag{48}$$

- A Bernoulli distribution $P(X)$ with the parameter θ, $P(X = x) = \theta^x (1 - \theta)^{(1-x)}$

$$E[X] = 1 \cdot \theta + 0 \cdot (1 - \theta) = \theta \tag{49}$$

Variance

The variance of a random variable gives a measure of how much the values of this random variable vary

$$
\begin{aligned}
\mathrm{Var}[X] &= E\left[(X - E[X])^2\right] \\
&= E\left[X^2 - 2XE[X] + E[X]^2\right] \\
&= E\left[X^2\right] - 2E[X]E[X] + E[X]^2 \\
&= E\left[X^2\right] - E[X]^2 \qquad (50)
\end{aligned}
$$

Variance: Example

A Bernoulli distribution $P(X)$ with the parameter θ,
$P(X = x) = \theta^x (1 - \theta)^{(1-x)}$

$$\mathrm{Var}[X] = E\left[X^2\right] - E\left[X\right]^2 = p - p^2 \tag{51}$$

Exercise: Compute the mean and variance of a categorical distribution

Statistical Estimation

Statistics is, in a certain sense, the inverse of probability theory.

- Observed: values of random variables
- Unknown: the model
- Task: infer the model from the observed data

Likelihood-based Estimation

For a probability $P(X;\theta)$ with θ as the unknown parameter, likelihood-based estimation with observations $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ requires two steps

1. Define a likelihood function with observations
2. Optimize the likelihood function to estimate θ

Likelihood Function

The likelihood function of θ is defined as

$$L(\theta) = \prod_{i=1}^{n} P(x^{(i)}; \theta) \tag{52}$$

Alternatively, we often use the log-likelihood function to avoid the numerical issues

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{n} \log P(x^{(i)}; \theta) \end{aligned} \tag{53}$$

Maximum Likelihood Estimation

Maximum Likelihood Estimation: a method of estimating the parameter by maximizing the (log-)likelihood function

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \ell(\theta) \tag{54}$$

Usually, this can be done with the following equation

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log P(x^{(i)}; \theta)}{\partial \theta} = 0 \tag{55}$$

Example: Bernoulli Distribution

Consider a Bernoulli distribution $P(X;\theta)$ with the parameter $\theta = P(X=1;\theta)$ unknown

$$P(X = x;\theta) = \theta^{x}(1-\theta)^{(1-x)} \tag{56}$$

Example: Bernoulli Distribution

Consider a Bernoulli distribution $P(X;\theta)$ with the parameter $\theta = P(X=1;\theta)$ unknown

$$P(X=x;\theta) = \theta^{x}(1-\theta)^{(1-x)} \tag{56}$$

With n observations $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$, the likelihood function is

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \log P(x^{(i)};\theta) \\ &= \sum_{i=1}^{n} \{x^{(i)} \log\theta + (1-x^{(i)})\log(1-\theta)\} \end{aligned} \tag{57}$$

Example: Bernoulli Distribution (II)

The derivative with respect to θ

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} \{ \frac{x^{(i)}}{\theta} - \frac{1 - x^{(i)}}{1 - \theta} \} \tag{58}$$

Example: Bernoulli Distribution (II)

The derivative with respect to θ

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{i=1}^{n} \{ \frac{x^{(i)}}{\theta} - \frac{1 - x^{(i)}}{1 - \theta} \} \tag{58}$$

Let $\frac{\partial \ell(\theta)}{\partial \theta} = 0$, we have

$$\theta = \frac{\sum_{i=1}^{n} x^{(i)}}{n} \tag{59}$$

Example: Bernoulli Distribution (III)

Assume the $n = 7$ observations are

$$\{0, 1, 1, 0, 0, 1, 0\},$$

then

$$\theta = \frac{3}{7} \tag{60}$$

Further Reference [Murphy, 2012, Chap 5 & 6]

Example: Bernoulli Distribution (III)

Assume the $n = 7$ observations are

$$\{0, 1, 1, 0, 0, 1, 0\},$$

then

$$\theta = \frac{3}{7} \tag{60}$$

Likelihood Principle: With $\boldsymbol{x}$ observed, all relevant information of inferring θ is contained in the likelihood function.

Further Reference [Murphy, 2012, Chap 5 & 6]

Reference

Eisenstein, J. (2018).
Natural Language Processing.
MIT Press.

Kolter, Z. and Do, C. (2015).
Linear algebra review and reference.

Murphy, K. P. (2012).
Machine learning: a probabilistic perspective.
MIT press.