CS 4774 Machine Learning

Introduction

Yangfeng Ji

Information and Language Processing Lab
Department of Computer Science
University of Virginia

Course Instructor and Teaching Assistants

- Instructor:
  - Yangfeng Ji
  - Office: Rice 510
- Teaching assistants
  - Oishee Hoque
  - Aparna Kishore
  - Nusrat Mozumder
  - Sanchit Sinha
  - Dane Williamson

Office hours will be released on the course webpage.

Prerequisites

- Programming and Algorithm
  - CS 2150 or CS 3100 with a grade of C- or better

Prerequisites

- Programming and Algorithm
  - CS 2150 or CS 3100 with a grade of C- or better
- Linear Algebra
  - Math 3350 or APMA 3080 or equivalent

Prerequisites

- Programming and Algorithm
  - CS 2150 or CS 3100 with a grade of C- or better
- Linear Algebra
  - Math 3350 or APMA 3080 or equivalent
- Probability and Statistics
  - APMA 3100, APMA 3110, MATH 3100, or equivalent

Goal

The survey results (by Jan. 18, 12 PM)

Attempts: 83 out of 83

What is the purpose of taking this course?

Please select the answers that are aligned with your expectation.

My research needs machine learning	6 respondents	7 %
To learn the basic idea of machine learning	65 respondents	**78 %**
To have a machine learning course on my transcript	26 respondents	31 %
To learn machine learning tools, e.g., PyTorch, Sklearn	61 respondents	73 %
To learn how to use machine learning solving problems	64 respondents	77 %
Machine learning is a hot topic	52 respondents	63 %
No Answer	1 respondent	1 %

Outline

This course will cover the basic materials on the following topics

1. Introduction to learning theory
2. Linear classification and regression
3. Model selection and validation
4. Boosting
5. Optimization methods
6. Neural networks (e.g., CNN, RNN, Auto-encoders, Transformers)

Outline (II)

The following topics will not be the emphasis of this course

- Statistical modeling
  - e.g., parameter estimation, Bayesian statistics, graphical models

Outline (II)

The following topics will not be the emphasis of this course

- Statistical modeling
  - e.g., parameter estimation, Bayesian statistics, graphical models
- Machine learning engineering
  - e.g., how to implement a classifier from end to end
  - although, we will provide some demo code for illustration purposes

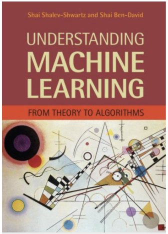
Outline (II)

The following topics will not be the emphasis of this course

- Statistical modeling
  - e.g., parameter estimation, Bayesian statistics, graphical models
- Machine learning engineering
  - e.g., how to implement a classifier from end to end
  - although, we will provide some demo code for illustration purposes
- Advanced topics in machine learning
  - e.g., reinforcement learning, active learning, semi-supervised learning, online learning

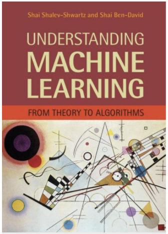
Textbooks

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

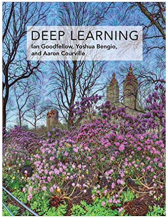


Textbooks

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



- Goodfellow, Bengio, and Courville. *Deep Learning*. 2016



Reference Books

For students looking for additional reading materials

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

Homework and Grading Policy

- Homeworks (70%)
  - Five homework assignments, 14 points each

Homework and Grading Policy

- Homeworks (70%)
  - Five homework assignments, 14 points each
- In-class Quiz (10%)
  - For the instructor to get a better understanding of students' feedback on the lectures
  - 1 point each

Homework and Grading Policy

- Homeworks (70%)
  - Five homework assignments, 14 points each
- In-class Quiz (10%)
  - For the instructor to get a better understanding of students' feedback on the lectures
  - 1 point each
- Final project (20%)
  - There are some pre-defined problems with datasets provided
  - Students will team up (3 – 4 students per group) to solve one problem

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
- Submission will not be accepted if more than 72 hours late
- Make sure not submit wrong files
  - it is students responsbility to make sure they submit the right and complete files for each homework
- It is usually better if students just turn in what they have in time

Violation of the Honor Code

Plagiarism, examples are

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

Webpages

- Course webpage

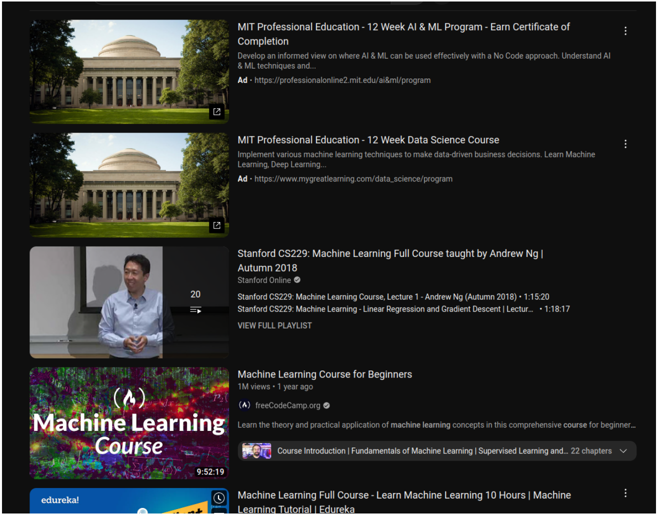
  `http://yangfengji.net/uva-ml-undergrad/`

  which contains all the information you need about this course.

- Canvas
  - For homework releasing and grading
  - For announcement, online QA, discussion, etc.

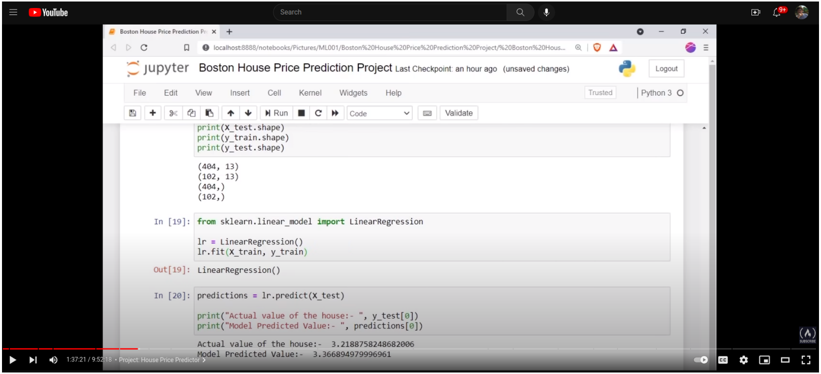
Why Taking this Course?

Machine Learning Courses



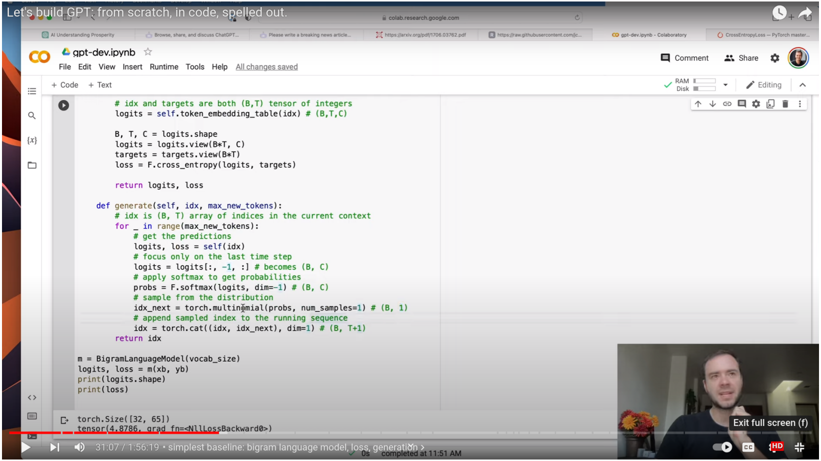
An Example Course

Building logistic regression classifier

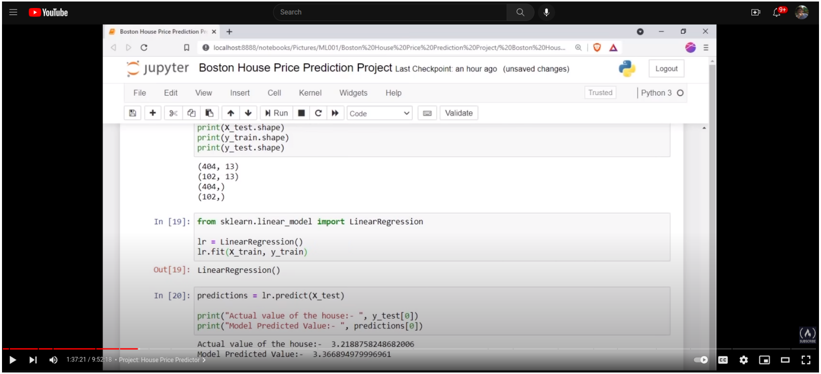


Another Example

Building a GPT model with Transformer



What is Missing?



What is this?

sklearn.linear_model.LogisticRegression

class sklearn.linear_model.**LogisticRegression**(*penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None*) [source]

Logistic Regression (aka logit, MaxEnt) classifier.

In the multiclass case, the training algorithm uses the one-vs-rest (OvR) scheme if the 'multi_class' option is set to 'ovr', and uses the cross-entropy loss if the 'multi_class' option is set to 'multinomial'. (Currently the 'multinomial' option is supported only by the 'lbfgs', 'sag', 'saga' and 'newton-cg' solvers.)

This class implements regularized logistic regression using the 'liblinear' library, 'newton-cg', 'sag', 'saga' and 'lbfgs' solvers. **Note that regularization is applied by default**. It can handle both dense and sparse input. Use C-ordered arrays or CSR matrices containing 64-bit floats for optimal performance; any other input format will be converted (and copied).

The 'newton-cg', 'sag', and 'lbfgs' solvers support only L2 regularization with primal formulation, or no regularization. The 'liblinear' solver supports both L1 and L2 regularization, with a dual formulation only for the L2 penalty. The Elastic-Net regularization is only supported by the 'saga' solver.

- What's the definition of this classifier?
- What if it does not work?
- What are its limitations?

What is this?

sklearn.linear_model.LogisticRegression

class `sklearn.linear_model.`**`LogisticRegression`**(*penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None*) [source]

Logistic Regression (aka logit, MaxEnt) classifier.

In the multiclass case, the training algorithm uses the one-vs-rest (OvR) scheme if the 'multi_class' option is set to 'ovr', and uses the cross-entropy loss if the 'multi_class' option is set to 'multinomial'. (Currently the 'multinomial' option is supported only by the 'lbfgs', 'sag', 'saga' and 'newton-cg' solvers.)

This class implements regularized logistic regression using the 'liblinear' library, 'newton-cg', 'sag', 'saga' and 'lbfgs' solvers. **Note that regularization is applied by default**. It can handle both dense and sparse input. Use C-ordered arrays or CSR matrices containing 64-bit floats for optimal performance; any other input format will be converted (and copied).

The 'newton-cg', 'sag', and 'lbfgs' solvers support only L2 regularization with primal formulation, or no regularization. The 'liblinear' solver supports both L1 and L2 regularization, with a dual formulation only for the L2 penalty. The Elastic-Net regularization is only supported by the 'saga' solver.

- What's the definition of this classifier?
- What if it does not work?
- What are its limitations?

In fact, if you explain how these parameters work and their effects, you can skip at least one third of the class lectures.

Mathematical/Statistical Background

To understanding machine learning, we need some mathematical and statistical knowledge

Attempts: 84 out of 84

The course materials also contain a large amount of mathematical/statistical stuff.

Please select the following answer that gives the closest description of how comfortable when you read mathematics.

No way for me to read any mathematics	1 respondent	1 %
Not a fan of mathematics, but I can handle it	31 respondents	37 %
I will try to avoid it, unless I have to	9 respondents	11 %
I feel comfortable of reading mathematics	43 respondents	**51 %** ✓

Now, let's have some fun!

Warning: you will see lots of mathematical notations.

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

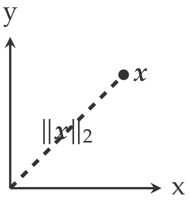
Vector Norms

- A norm of a vector $\|\boldsymbol{x}\|$ is informally a measure of the "length" of the vector.
- Formally, a norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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

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



✎ *Question for Homework*: prove ℓ_2 norm satisfies all four properties

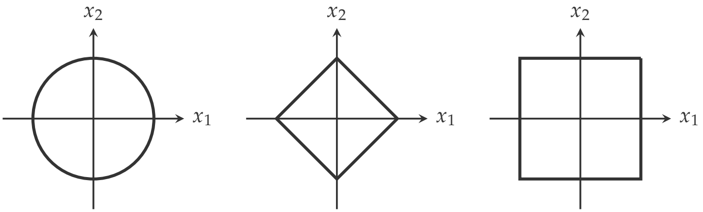
ℓ_1 Norms

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

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

Plots

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)

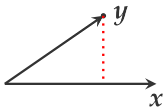
Dot Product

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

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

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

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

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

Matrix-Vector Multiplication

Given a matrix $\boldsymbol{A}$ and a vector $\boldsymbol{x}$, their multiplication is equivalent to performing a linear transformation on $\boldsymbol{x}$

$$\boldsymbol{A}\boldsymbol{x} \tag{9}$$

For example, consider the following matrix

$$\boldsymbol{A} = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix} \tag{10}$$

and three vectors

- $\boldsymbol{x}_1^\top = [1, 2]$
- $\boldsymbol{x}_2^\top = [2, 4]$
- $\boldsymbol{x}_3^\top = [3, 6]$

Matrix-Vector Multiplication

Given a matrix $\boldsymbol{A}$ and a vector $\boldsymbol{x}$, their multiplication is equivalent to performing a linear transformation on $\boldsymbol{x}$

$$\boldsymbol{A}\boldsymbol{x} \tag{9}$$

For example, consider the following matrix

$$\boldsymbol{A} = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix} \tag{10}$$

and three vectors

- $\boldsymbol{x}_1^\top = [1, 2]$
- $\boldsymbol{x}_2^\top = [2, 4]$
- $\boldsymbol{x}_3^\top = [3, 6]$

➥ *This is also what the function torch.nn.Linear means*

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.

$$\mathbf{I} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \tag{11}$$

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.

$$\mathbf{I} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \tag{11}$$

- 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.

$$\mathbf{D} = \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{bmatrix} \tag{12}$$

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

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} = \boldsymbol{A}\boldsymbol{A}^{-1} \tag{13}$$

- 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 $\boldsymbol{x} = \mathbf{A}^{-1}\boldsymbol{b}$

Inverse (II)

- In matrix-vector multiplication, an inverse matrix $\boldsymbol{A}^{-1}$ will reverse the linear transformation performed by $\boldsymbol{A}$

$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x} \tag{14}$$

Inverse (II)

- In matrix-vector multiplication, an inverse matrix $\boldsymbol{A}^{-1}$ will reverse the linear transformation performed by $\boldsymbol{A}$

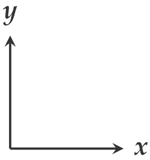
$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x} \tag{14}$$

- When a matrix $\boldsymbol{A}$ is not invertible, it means its linear transformation is not reversible

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{15}$$

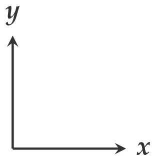
Orthogonal Matrices

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



Orthogonal Matrices

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



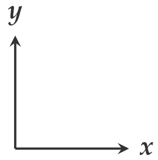
- A square matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ is orthogonal, if all its columns are orthogonal to each other *and* normalized (orthonormal)

$$\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = 0, \|\boldsymbol{u}_i\| = 1, \|\boldsymbol{u}_j\| = 1 \tag{16}$$

for $i, j \in [n]$ and $i \neq j$

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)

$$\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle = 0, \|\boldsymbol{u}_i\| = 1, \|\boldsymbol{u}_j\| = 1 \tag{16}$$

  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 $\mathbf{U}^{-1} = \mathbf{U}^\top$

A Special Case

Consider a matrix $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{17}$$

For any $\boldsymbol{x}^\top = [x_1, x_2]$, we have

$$\boldsymbol{A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} \tag{18}$$

A Special Case

Consider a matrix $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{17}$$

For any $\boldsymbol{x}^\top = [x_1, x_2]$, we have

$$\boldsymbol{A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} \tag{18}$$

The is a reflection operation. Operations like this are popularly used in computer graphics.

Symmetric Matrices

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

$$\mathbf{A}^\top = \mathbf{A} \tag{19}$$

or, in other words,

$$a_{i,j} = a_{j,i} \quad \forall i, j \in [n] \tag{20}$$

Comments

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

$$\boldsymbol{x}^\top A \boldsymbol{y} \tag{21}$$

gives each dimension a different weight (importance) when computing the similarity between $\boldsymbol{x}$ and $\boldsymbol{y}$

Eigen Decomposition

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

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \tag{22}$$

with

- $\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$ as a diagonal matrix
- $\mathbf{Q}$ is an orthogonal matrix

Eigen Decomposition

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

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top \tag{22}$$

with

- $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$ as a diagonal matrix
- $\mathbf{Q}$ is an orthogonal matrix
- Consider the similarity measurement

$$\boldsymbol{x}^\top \mathbf{A}\boldsymbol{y} = \boldsymbol{x}^\top \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top \boldsymbol{y} = (\mathbf{U}^\top \boldsymbol{x})^\top \boldsymbol{\Lambda}\mathbf{U}^\top \boldsymbol{y} \tag{23}$$

Eigen Decomposition

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

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \tag{22}$$

with

- $\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$ as a diagonal matrix
- $\mathbf{Q}$ is an orthogonal matrix
- Consider the similarity measurement

$$\boldsymbol{x}^\top \mathbf{A} \boldsymbol{y} = \boldsymbol{x}^\top \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \boldsymbol{y} = (\mathbf{U}^\top \boldsymbol{x})^\top \mathbf{\Lambda}\mathbf{U}^\top \boldsymbol{y} \tag{23}$$

✎ *Question for Homework*: if a symmetric matrix $\mathbf{A}$ is invertible, show $\mathbf{A}^{-1} = \mathbf{U}\Lambda^{-1}\mathbf{U}^\top$ with $\Lambda^{-1} = \mathrm{diag}(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n})$

Symmetric Positive Semidefinite Matrices

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

$$\boldsymbol{x}^\top \mathbf{P} \boldsymbol{x} \geq 0 \tag{24}$$

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

$$\boldsymbol{x}^\top \mathbf{P} \boldsymbol{x} \geq 0 \tag{24}$$

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

Eigen decomposition of $\mathbf{P}$ as

$$\mathbf{P} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^\top \tag{25}$$

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

$$\lambda_i \geq 0 \tag{26}$$

Symmetric Positive Definite Matrices

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

$$\boldsymbol{x}^\top \mathbf{P} \boldsymbol{x} > 0 \tag{27}$$

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

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

$$\lambda_i > 0 \tag{28}$$

Symmetric Positive Definite Matrices

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

$$\boldsymbol{x}^\top \mathbf{P} \boldsymbol{x} > 0 \tag{27}$$

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

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

$$\lambda_i > 0 \tag{28}$$

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

Review

The identity matrix $\mathbf{I}$ is

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

Review

The identity matrix $\mathbf{I}$ is

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

Review of 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

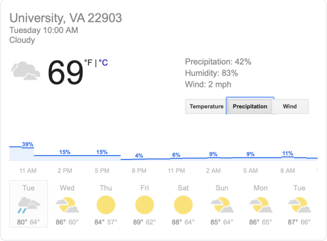
- If we flip the coin many times, we expect it to land heads about 52% times

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

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

Bayesian Interpretation

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

$$P(X = 1) = 0.52 \tag{29}$$

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

$$P(Y = k) = \binom{n}{k}\theta^k(1 - \theta)^{n-k} \tag{30}$$

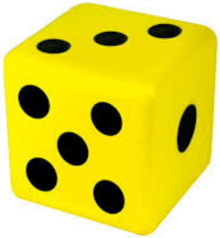
where

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

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

$$P(\boldsymbol{X} = \boldsymbol{x}) = \prod_{k=1}^{6} (\theta_k)^{x_k} \tag{31}$$

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

$$P(\boldsymbol{X} = \boldsymbol{x}) = \binom{n}{x_1 \cdots x_K} \prod_{k=1}^{K} \theta_k^{x_k} \tag{32}$$

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:

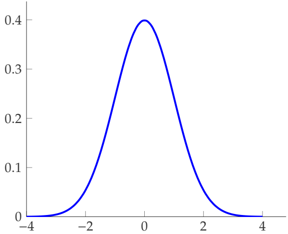
$$\sum_{k=1}^{K} x_k = n$$

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

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \tag{33}$$

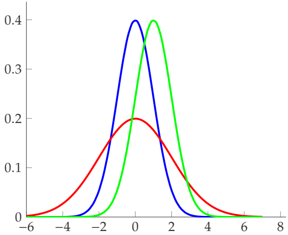
- μ: mean
- σ^2: variance
- Probability of $X \in [a, b]$: $P(a \leq X \leq b) = \int_a^b f(x)dx$



Gaussian Distribution (II)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right) \tag{34}$$

There examples of Gaussian distributions

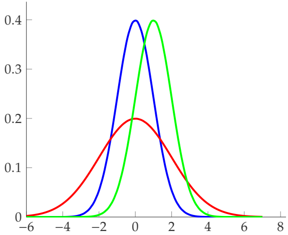


- Blue: $\mathcal{N}(0, 1)$ (standard normal distribution)
- Red: $\mathcal{N}(0, 2)$
- Green: $\mathcal{N}(1, 1)$

Gaussian Distribution (II)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{34}$$

There examples of Gaussian distributions



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

✎ *Question for Homework*: describe a random event with the probabilistic language

Probability of Two Random Variables

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

$$P(X, Y) \tag{35}$$

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,

$$P(X, Y) = P(X) \cdot P(Y) \tag{36}$$

Marginal distributions

$$P(X) = \sum_{Y} P(X, Y) \tag{37}$$

$$P(Y) = \sum_{X} P(X, Y) \tag{38}$$

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,

$$P(X, Y) = P(X) \cdot P(Y) \tag{36}$$

Marginal distributions

$$P(X) = \sum_{Y} P(X, Y) \tag{37}$$

$$P(Y) = \sum_{X} P(X, Y) \tag{38}$$

- 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

$$P(Y \mid X) = \frac{P(X, Y)}{P(X)} \tag{39}$$

Example: document classification

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

A special case: if X and Y are independent

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

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

Conditional Probability

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

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

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

Conditional Probability

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

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

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

Conditional Probability

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

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

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

✎ *Question for Homework*: compute conditional probability from a given probabilistic table

Multivariate Gaussian

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

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right) \tag{41}$$

where

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

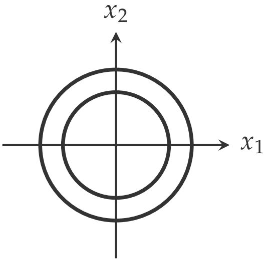
Covariance Matrix Σ

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

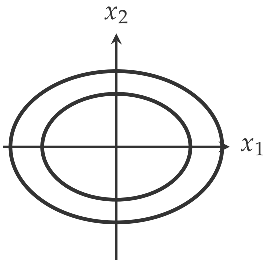
$$f(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right) \tag{42}$$

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

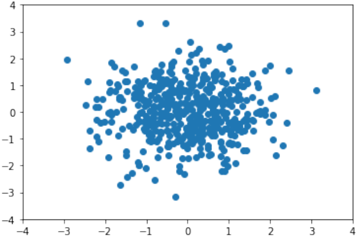
$\boldsymbol{\Sigma} = \mathbf{I}$



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



Sampling from Gaussians



(a) (b)

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

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

✎ *Question for Homework*: sample from a Gaussian distribution with a pre-defined mean and variance

Thank You!