Introduction

STATS 303 Statistical Machine Learning

Spring 2022

Lecture 1

syllabus

instructor's information

instructor:

- Dongmian Zou, Assistant Professor of Data Science
  - email: dongmian.zou@dukekunshan.edu.cn
  - office hour: Tuesday 11:15am – 12:15pm; Thursday 3pm – 4pm; or by appointment (let me know beforehand if you want to join a zoom room)

TA:

- recitation: Eric Qu
  - email: zhonghang.qu@dukekunshan.edu.cn
- homework: Xue Chen
  - email: xue.chen240@dukekunshan.edu.cn

What will I do in this course?

- lectures + recitation
- homework assignments
- presentation
- midterm and final exams (open-book)

lectures and recitation

- synchronous meeting time:
  - lectures: Monday – Thursday 8:30am – 9:45am
  - recitation: Thursday 7pm – 8pm
  - At the beginning of the course, you will form groups (of 4 ~ 5). Each week, there is a worksheet for your group to work together on. During the recitation sessions, Eric will lead the discussion.

presentation

- You will work in your group to explore a topic of interest that is not covered in class.
- During the last week of lectures, you will give a presentation on your discovery.
- The detailed rubric for this is on Sakai.

presentation

- The topic must be relevant to statistical machine learning and not covered in STATS302/303.
- You can either choose a topic and collect relevant materials, e.g.
    - conditional random field
    - upper confidence bound (UCB) algorithm in reinforcement learning
    - concentration inequalities
    - genetic algorithms
    - differential privacy

presentation

- Or choose a paper in statistical machine learning and present the technical contents, e.g.
  - Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005, August). Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning (ICML)* (pp. 89-96).
  - Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems (Neurlips)*, 20.
- If you choose to work with a paper, make sure that it is understandable (*oldies and goldies* may work well).

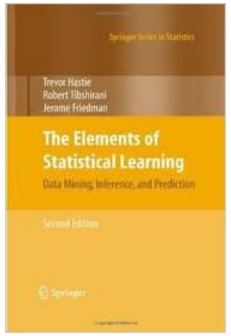
grades

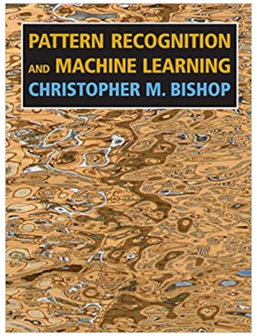
activity	points	comments
homework	20%	submit on Sakai; 6 in total, lowest score dropped
presentation	10%	submit slides on Sakai; deliver during the last week
midterm	30%	open-book
final	40%	open-book

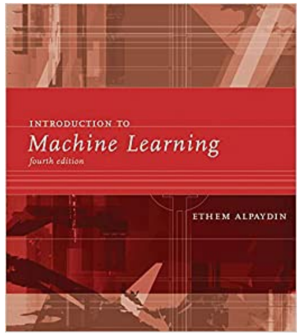
Please refer to the following scale for your grading.

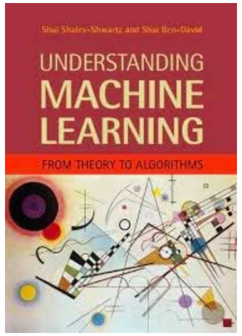
A+= 98% - 100%; **A** = 97% - 93%; **A-** = 90% - 92%; **B+** = 87% - 89%; **B** = 83% - 86%; **B-** = 80% - 82%; **C+** = 77% - 79%; **C** = 73% - 76%; **C-** = 70% - 72%; **D+** = 67% - 69%; **D** = 63% - 66%; **D-** = 60% - 62%; **F** = 59% and below

textbooks





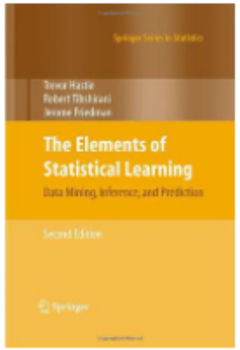




- Lectures are self-contained and slides will be posted on Sakai (slides will be updated after each class).
- Lectures are based mainly on the following books:
  - **Elements of statistical learning [HaTF]** by Hastie, Tibshirani and Friedman
    - available at the official webpage: https://web.stanford.edu/~hastie/ElemStatLearn/
  - **Pattern recognition and machine learning [Bi]** by Bishop
    - available at Microsoft webpage: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
  - **Introduction to Machine Learning [Al]** by Alpaydin
    - available from the Duke library: https://find.library.duke.edu/catalog/DUKE007630227
  - **Understanding Machine Learning: From Theory to Algorithms [S-S]** by Shalev-Shwartz and Ben-David
    - available at the official webpage: https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/copy.html

textbooks

The Elements of Statistical Learning

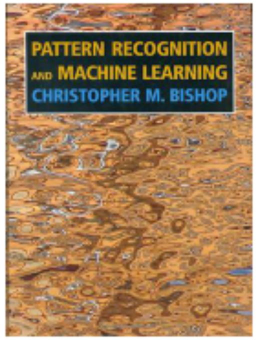


Author : Trevor Hastie / Robert Tibshirani / Jerome Friedman
Publisher: Springer
subtitle: the Data Mining, Inference, and Prediction, Second Edition
Published: 2009-10-1
Pages: 745
Price: GBP 62.99
Binding: Hardcover
Series: Springer Series in Statistics

Douban score
9.4
677 Ratings

5 stars	78.9
4 stars	18.3%
3 stars	2.4%
2 stars	0.1%
1 star	0.3%

Pattern Recognition and Machine Learning



Author : Christopher Bishop
Publisher: Springer
Published: 2007-10-1
Pages: 738
Price: USD 94.95
Binding: Hardcover
ISBN: 9,780,387,310,732

Douban score
9.5
1351 Ratings

5 stars	82.0%
4 stars	15.5%
3 stars	2.4%
2 stars	0.1%
1 star	0.1%

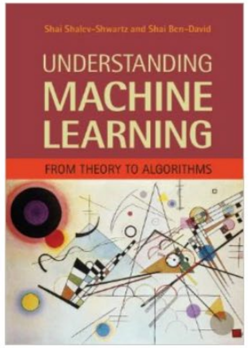
Introduction to Machine Learning

Author : Ethen Alpaydin
Press: Machinery Industry Press
Original Title: Introduction to Machine Learning translator : Fan Ming / Zan Hongying / cow Chang Yong Published: 2009-6 Pages: 272 Price: 39.00 yuan Binding: Paperback
Series: Computer Science Series ISBN : 9787111265245

Douban score
7.2
92 Ratings

5 stars	17.4%
4 stars	48.
3 stars	23.9%
2 stars	8.7%
1 star	1.1%

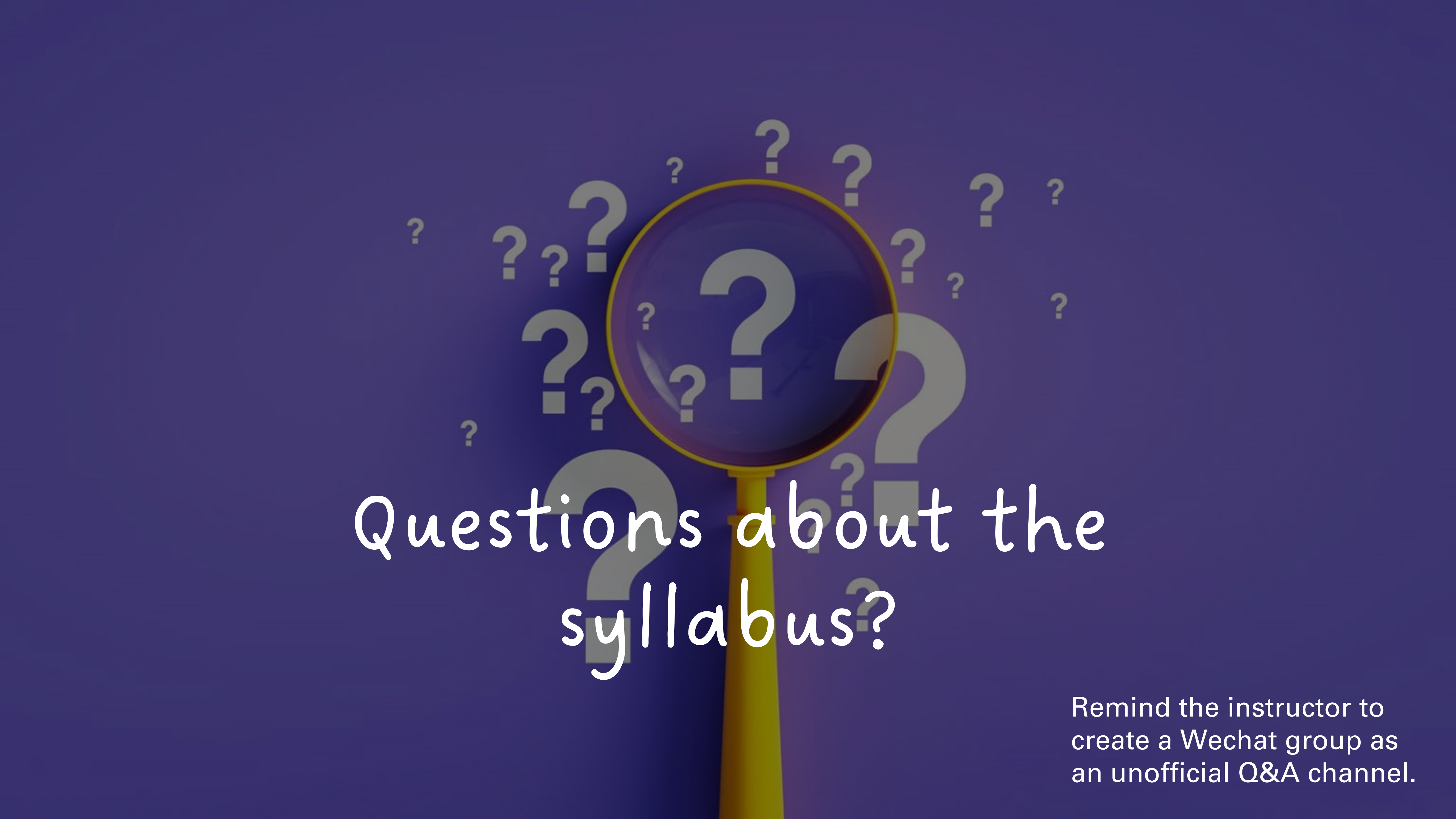
Understanding Machine Learning



Author : Shai Shalev-Shwartz / Shai Ben-David
Publisher: Cambridge University Press
Subtitle: From Theory to Algorithms
Publication year: 2014
Pages: 424
Price: USD 48.51Finishing
: Hardcover
ISBN: 9781107057135

Douban score
8.5
66 ratings

5 stars	62.1%
4 stars	28.8%
3 stars	7.6%
2 stars	1.5%
1 star	0.0%



Questions about the syllabus?

Remind the instructor to create a Wechat group as an unofficial Q&A channel.

What is statistical machine learning?

a little bit history

1950s
Statistical methods are used in machine learning

1960s
Bayesian methods are introduced

1970s
1st “AI Winter”

1980s
Rediscovery of back-propagation

Late 1980s
2nd “AI Winter”

1990s
Data-driven approaches; SVM and RNN become popular

2000s
Clustering and kernel methods become popular

2010s
Deep learning

2020s
???

Statistical Learning
emphasizes statistical principles +
mathematical frameworks for making inference

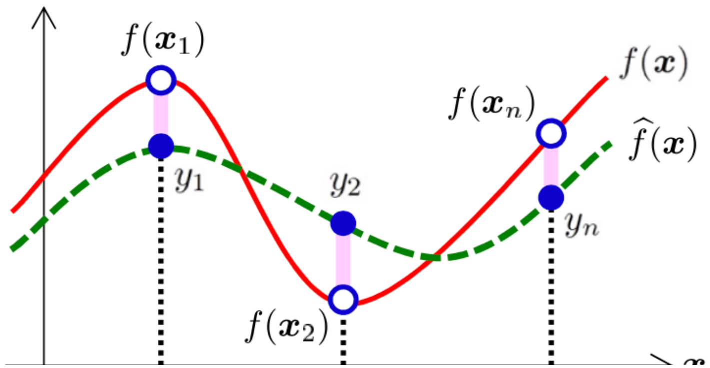
=

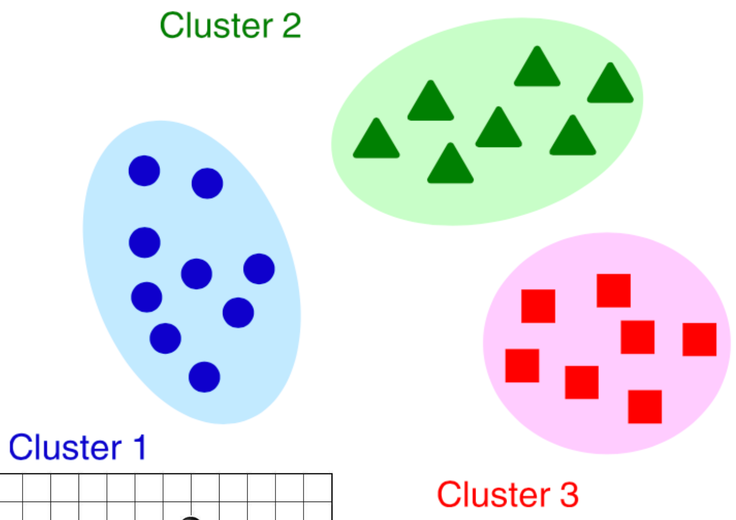
machine learning
emphasizes computer algorithms +
efficient implementation

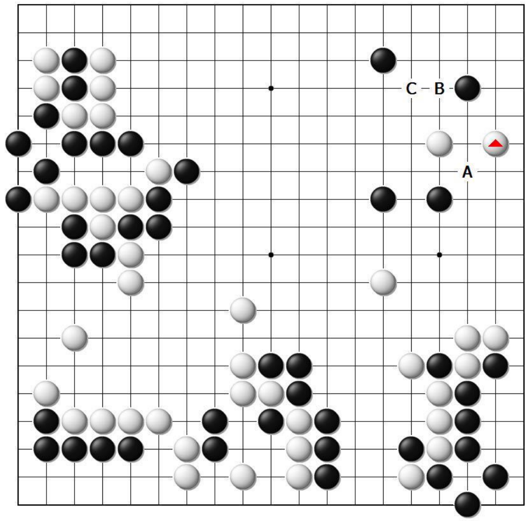
data science
suggested by Michael Jordan to call the overall field

categories of machine learning

- supervised learning
  - learn through Q&A from a supervisor (objective: *generalization*)
  - e.g. regression, classification
- unsupervised learning
  - learn by himself/herself
  - e.g. clustering, outlier detection
- reinforcement learning
  - the supervisor does not directly give answers to the student’s questions but gives feedbacks
  - e.g. computer games, robots







Bayesian decision theory

unobservable and observable variables

- Suppose we toss a coin; the outcome will be either a head (H) or a tail (T).
- If we have extra knowledge, e.g., the exact composition of the coin, its initial position, the force applied to the coin, and forth, the exact outcome of the toss could be predicted.
- The extra pieces of knowledge that we do not have access to are named the unobservable variables, or latent variables.
- In the example of coin tossing, the only observable variable is the outcome (H or T).
- We have, in reality, $$\underbrace{x}_{\text{observable}} = \overbrace{f}^{\text{deterministic function}}(\underbrace{z}_{\text{unobservable}})$$

random variables

- We don't have access to the z, so we define the outcome X as a random variable drawn from a probability distribution $P(X = x)$ that specifies the process.
- In the coin tossing example, let $X = 1$ for (H) and $X = 0$ for (T)

$$P(X = 1) = p_0$$

$$P(X = 0) = 1 - P(X = 1) = 1 - p_0$$

samples

- If we don't know $P(X)$ and want to estimate this from a given sample, then we are in the realm of statistics
- We have a **sample** $\chi = \{x_n\}_{n=1}^{N}$ drawn from the probability distribution of the observables $p(x)$

- Aim: build an approximator $\hat{p}(x)$ using the sample χ
- In the coin tossing example,

$$\hat{p}_0 = \frac{\#\{\text{tosses with outcome } (H)\}}{\#\{\text{tosses}\}}$$

Classification

- Suppose we work in a bank, and would like to learn the classes "high-risk customer" [C=1] and "low-risk customer" [C=0].

- We decide there are two pieces of information available:

  X_1 : yearly income

  X_2 : savings

- If we know $P(C \mid X_1, X_2)$, when a new application arrives with $X_1 = x_1$, $X_2 = x_2$, we can choose

$$C = 1 \quad \text{if} \quad P(C=1 \mid X_1 = x_1, X_2 = x_2) > 0.5$$

$$C = 0 \quad \text{otherwise}$$

classification

- Let $\boldsymbol{x} = [x_1, x_2]^{\mathrm{T}}$ .The problem is to calculate $P(C|\boldsymbol{x})$.
- By Bayes' Rule,

$$P(C|\boldsymbol{x}) = \frac{P(C)p(\boldsymbol{x}|C)}{p(\boldsymbol{x})}$$

- That is,

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

classification in general

- In general, we have K classes (mutually exclusive, and exhaustive): $C_1, C_2, \cdots, C_K$
- We have $P(C_i) \geq 0$ and $\sum_{i=1}^{K} P(C_i) = 1$

$p(x) = \sum_{k=1}^{K} p(x, C_k)$

$$P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)} = \frac{p(x|C_i)P(C_i)}{\sum_{k=1}^{K} p(x|C_k)P(C_k)}$$

- **Bayes' classifier**: choose the class with the highest posterior probability:

$$\text{choose } C_i \text{ if } P(C_i|x) = \max_{k=1,\cdots,K} P(C_k|x)$$

loss and risk

- Define
  - action α_i as the decision to assign the input to class C_i
  - λ_{ik} as the loss incurred for taking α_i when the input actually belongs to C_k (if we allow abuse of notation, we can say $\boldsymbol{x} \in C_k$).
- Then the **expected risk** for taking α_i is

$$R(\alpha_i|\boldsymbol{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|\boldsymbol{x})$$

loss and risk

- $R(\alpha_i|\boldsymbol{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|\boldsymbol{x})$
- In the special case of **0/1 loss**, where $\lambda_{ik} = \begin{cases} 0 & \text{if} \quad i = k \\ 1 & \text{if} \quad i \neq k \end{cases}$
- $R(\alpha_i|\boldsymbol{x}) = \sum_{k \neq i} P(C_k|\boldsymbol{x}) = 1 - P(C_i|\boldsymbol{x})$

reject

- In the above, we already have actions α_i as the decision to assign the input to class C_i , $i = 1,2,\cdots,K$

- Let's define an additional action of **reject** (not making any decision, indecisive): α_{K+1}

- By modifying the 0/1 loss, a possible loss function is

$$\lambda_{ik} = \begin{cases} 0 & \text{if} \quad i = k \\ 1 & \text{if} \quad i \in [K] - \{k\} \\ \lambda & \text{if} \quad i = K+1 \end{cases} = \begin{cases} 0 & \text{if} \quad i = k \\ \lambda & \text{if} \quad i = K+1 \\ 1 & \text{otherwise} \end{cases}$$

Here $[K] = \{1, 2, \cdots, K\}$

reject

- The risk of reject is $R(\alpha_{K+1}|\boldsymbol{x}) = \sum_{k=1}^{K} \lambda P(C_k|\boldsymbol{x}) = \lambda$
- The risk of choosing C_i is $1 - P(C_i|\boldsymbol{x})$

reject

- The optimal decision rule:

  - Choose C_i if

    (1) $R(\alpha_i|\boldsymbol{x}) < R(\alpha_k|\boldsymbol{x})$ for all $k \neq i$ and

    (2) $R(\alpha_i|\boldsymbol{x}) < R(\alpha_{K+1}|\boldsymbol{x})$

  - Reject if

    $R(\alpha_{K+1}|\boldsymbol{x}) < R(\alpha_i|\boldsymbol{x})$ for all i

Questions?

Reference

- *Introduction:*
  - *[Al] Ch.1*
  - *[HaTF] Ch.1*
- *Bayesian decision theory:*
  - *[Al] Ch.3.1-3.4*
  - *[HaTF] Ch.2.4*
- *Maximum likelihood:*
  - *[Al] Ch.4.1-4.3*
  - *[Bi] Ch.2.4*
  - *[HaTF] Ch.2.6, 8.2.2*