Statistical and Computational Foundations of Machine Learning

Lecture 01

Dávid Pál

New York University

January 29, 2023

Last modified: 2023/01/07 10:13:54 -05'00'

Overview

1. Administrative stuff
2. Learning materials
3. Introduction to machine learning
4. Probability distributions
5. Statistical model

Administrative stuff

Contact information

- Course website: http://david.palenica.com/ml2023/
- Lectures: Sundays 5:00pm–7:30pm,
  Jacobs Academic Building, Room 473
- Office hours: Wednesdays 9:30am–11:00am over Zoom
- Email: david.pal@nyu.edu
- Ed Discussion Forum (for questions):
  https://edstem.org/us/courses/20105/discussion/
- Zoom: https://nyu.zoom.us/my/dapal
- No office at NYU.

Course style

- This is a theoretical course.
- We will state definitions and prove theorems.
- In homeworks and exams, you will calculate and prove theorems.
- You will not code.

This course is very hard. If you are struggling, reach out to me as soon as possible.

Prerequisites

- Probability theory
- Calculus
- Linear algebra
- Complexity theory is a plus

Grading

Homework assignments	40%
Midterm exam	30%
Final exam	40%
Total	110%

Tentative grades based on the last year

- A if 90% or more
- A- if $[80\%, 90\%)$
- B+ if $[70\%, 80\%)$
- B if $[60\%, 70\%)$
- B- if $[50\%, 60\%)$

Homeworks

- Homeworks will be posted on the course website.
- Your solutions:
  - Use LaTeX. Template is on the course website.
  - Email PDF file to david.pal@nyu.edu
    Subject: `Homework #X`
  - Hand-in printed solution in class.

	Posted on	**Due date**	**Graded by**
Homework #0	January 24	**January 31**	February 5
Homework #1	February 5	**February 19**	February 26
Homework #2	February 19	**March 5**	March 12
Homework #3	March 26	**April 9**	April 16
Homework #4	April 9	**April 23**	April 30

Homework policy

Collaboration policy

- Discussing problems with other students is allowed.
- Writing up of the solution must be done individually.
- Copied solutions are considered cheating.

Midterm and final exams

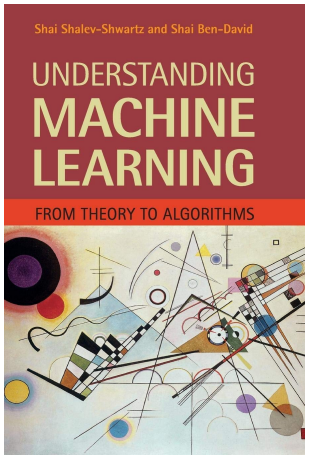
	Exam	**Graded by**
Midterm exam	**March 12**	March 26
Final exam	**May 14**	May 21

- Exams will be in person in class
- 2.5 hours (5:00pm – 7:30pm)
- Exam problems will be the same style as homework problems.
- Open book. Any paper materials are allowed (books, printed notes, handwritten notes, cheat-sheets, research papers).
- No electronics allowed (no computers, no phones, no tablets, no smart-watches)
- Communication with anybody else is prohibited.
- Sharing paper materials (books, notes) is prohibited.

Learning materials

Primary textbook

Understanding Machine Learning: From Theory to Algorithms
by Shai Shalev-Shwartz and Shai Ben-David
Cambridge University Press, 2014



Available online as a PDF, in NYU library, and on Amazon.

Shai Ben-David's video lectures on YouTube.

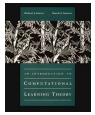
Today's lecture: Read chapters 1,2,3.

More books

- **Foundations of Machine Learning**
  by Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar
  MIT Press, 2012

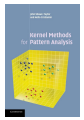
- **Neural Network Learning: Theoretical Foundations**
  by Martin Anthony and Peter L. Bartlett
  Cambridge University Press, 1999

- **An Introduction to Computational Learning Theory**
  by Michael J. Kearns and Umesh V. Vazirani
  MIT Press, 1994



- **Boosting: Foundations and Algorithms**
  by Robert E. Schapire and Yoav Freund
  MIT Press, 2014

- **Kernel Methods for Pattern Analysis**
  by John Shawe-Taylor and Nello Cristianini
  Cambridge University Press, 2004



Introduction to machine learning

Machine learning (ML)

Definition (Machine learning)

Machine learning (ML) is a discipline that designs and studies algorithms that "learn" i.e. improve through use of data.

As a scientific discipline, ML is a part of

1. Statistics
2. Optimization (a.k.a. operations research)
3. Computer science

ML is part of Artificial Intelligence (AI).
These days, ML is almost synonomous with AI.

Applications of machine learning

- Security
  - Spam filtering (email, discussion forums)
  - Credit card fraud detection
  - Network intrusion detection
- Content optimization
  - Web search
  - Recommendation systems (e.g. Netflix, Amazon, YouTube)
  - Online advertising
- Computer vision
  - Optical Character Recognition (OCR)
  - Face detection
  - Face recognotion
  - Object recognition
- Natural language processing
  - Speech-to-text
  - Translation
  - Chatbots

Types of machine learning

- Supervised vs. Unsupervised
  - Are there labels that need to predicted?
  - No obvious labels (clustering, topic modelling)?
- Active vs. Passive
  - Passive: Samples are collected at random.
  - Active: Samples are collected by careful selection.
- Online vs. Batch
  - Online: Learning and prediction is interleaved.
  - Batch: Collect data. Then learn.
- Adversarial vs. I.I.D. vs. Helpful
  - How are examples obtained?

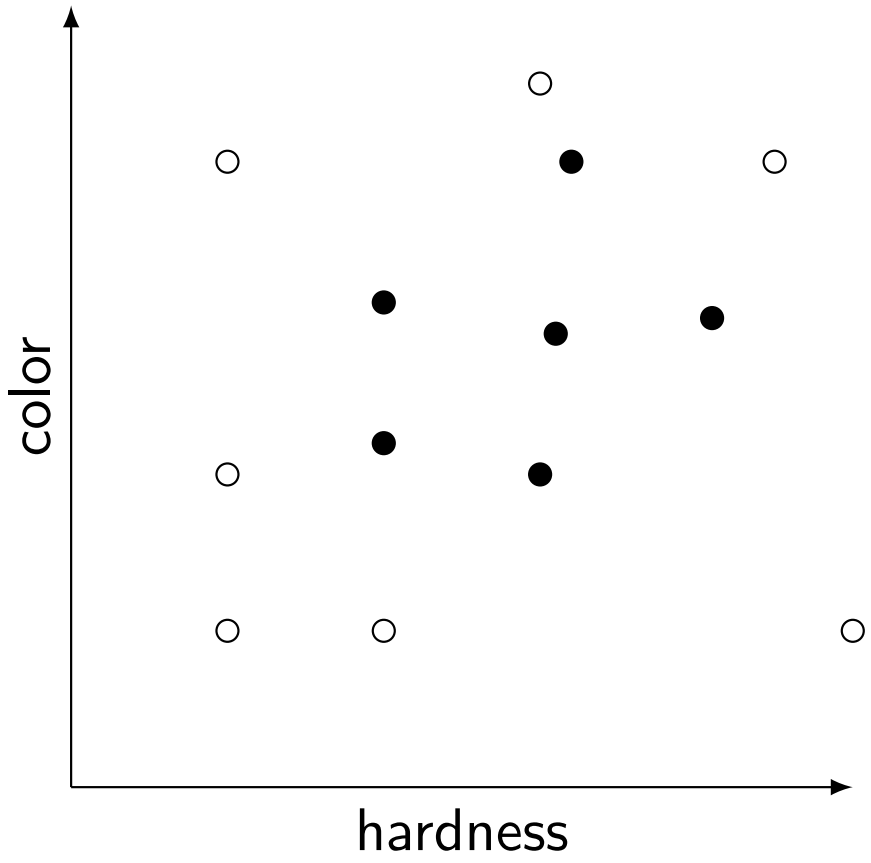
We will focus on supervised passive batch learning with i.i.d. data, with some excursions to adversarial online learning.
We will almost exclusively focuse on binary classification.

Which papaya's are tasty?

- Papaya is a fruit originating in Central America.
- It is grown in tropical regions all over the world (India, Australia, Hawaii, Florida).
- How do you know if a papaya fruit is tasty before buying it in a store?

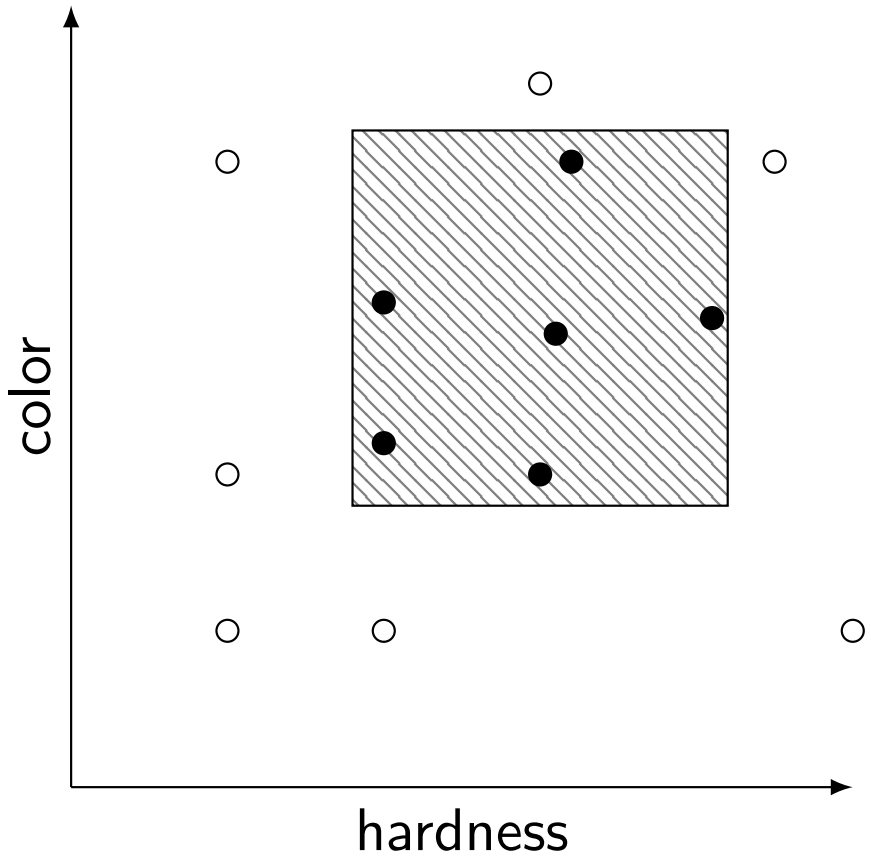
Which papaya's are tasty?

- There are two obvious features of the fruit:
  - color (from green, to yellow, to red, to brown)
  - hardness (from very soft to very hard)
- Suppose you collect a number of Papaya's and taste them all.



Which papaya's are tasty?

- There are two obvious features of the fruit:
  - color (from green, to yellow, to red, to brown)
  - hardness (from very soft to very hard)
- Suppose you collect a number of Papaya's and taste them all.



Probability distributions

Motivation

- We will use the concept of **probability distribution** a lot.
- Hopefully, you took some introductory probability theory course.
- However, you might have not seen formal definitions.
- We will go through an exceptionally brief introduction to **measure-theoretic probability**.
- Take a graduate course in probability theory to learn more.

σ-algebra

Definition (σ-algebra)

Let $\mathcal{X}$ be a set. A set system $\mathcal{F} \subseteq \mathcal{P}(X)$ is a *σ-algebra* on $\mathcal{X}$ if the following three conditions hold:

1. $\mathcal{X} \in \mathcal{F}$.
2. If $A \in \mathcal{F}$ then $\mathcal{X} \setminus A \in \mathcal{F}$.
3. If $A_1, A_2, \dots \in \mathcal{F}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

The pair $(\mathcal{X}, \mathcal{F})$ is called a *measurable space*.

Probability measure

Definition (Probability measure)

Let $\mathcal{X}$ be a set. Let $\mathcal{F}$ be σ-algebra on X. A *probability measure* on $(\mathcal{X}, \mathcal{F})$ is a function $P : \mathcal{F} \to \mathbb{R}$ that satisfies the following three conditions:

1. $P(A) \geq 0$ for all $A \in \mathcal{F}$
2. If $A_1, A_2, \dots \in \mathcal{F}$ are disjoint sets, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

3. $P(\mathcal{X}) = 1$

Probability space

Definition (Probability space)

A *probability space* is a triple $(\mathcal{X}, \mathcal{F}, P)$ where $(\mathcal{X}, \mathcal{F})$ is a measurable space and P is a probability measure on $(\mathcal{X}, \mathcal{F})$.

- We often omit $\mathcal{X}$ and/or $\mathcal{F}$ and talk about P only.
- We often refer to P as a *probability distribution* or *distribution*.
- Very often we use letter D instead of P.
  In measure theory, often the letter μ is used.

Example #1: Discrete probability measures

- $\mathcal{X}$ is finite or countable
- $\mathcal{F} = \mathcal{P}(\mathcal{X})$ i.e. $\mathcal{P}$ contains all subsets of $\mathcal{X}$.
- *Probability mass* function $p : \mathcal{X} \to [0, 1]$ such that

$$\sum_{x \in \mathcal{X}} p(x) = 1 \,.$$

- We define the probability measure as

$$P(A) = \sum_{x \in A} p(x) \qquad \text{for any } A \in \mathcal{F}.$$

Example #2: Continuous probability measures on $\mathbb{R}$

- $\mathcal{X} = \mathbb{R}$ i.e. set of real numbers
- $\mathcal{F}$ consists of so-called Borel sets[1].
- *Probability density function* function $p : \mathcal{X} \to [0, \infty)$ such that

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1\,.$$

- We define the probability measure as

$$P(A) = \int_A p(x)\,\mathrm{d}x \qquad \text{for any } A \in \mathcal{F}.$$

[1] We are not going define the Borel sets in this course.

Statistical model

Notation and definitions #1

- Non-empty set $\mathcal{X}$ called *domain* or *feature space*
- An element $x \in \mathcal{X}$ is called an *unlabeled example*
- A sequence of unlabeled examples is called a *unlabeled sample*

$$U = (x_1, x_2, \ldots, x_m) \in \mathcal{X}^* \,.$$

  where

$$\mathcal{X}^* = \bigcup_{m=0}^{\infty} \mathcal{X}^m$$

  is the set of all sequences of finite length.
- Set of *labels* $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{+1, -1\}$
- A pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is called a *labeled example*
- A sequence of labeled examples is called a *labeled sample*:

$$S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^*$$

  where $(\mathcal{X} \times \mathcal{Y})^*$ is the set of all sequences of finite length.

Notation and definitions #2

- A function $h : \mathcal{X} \to \mathcal{Y}$ is called a *classifier*.
  Many different names:
  - *classifier*
  - *predictor*
  - *hypothesis*
  - *concept*
  - *model*
- A classifier $h : \mathcal{X} \to \mathcal{Y}$ corresponds to the set $h^{-1}(1)$.
- Set all functions from $\mathcal{X}$ to $\mathcal{Y}$ is denoted $\mathcal{Y}^{\mathcal{X}}$.
- A *learning algorithm* is a function $A : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$.
  Many different names:
  - *learning algorithm*
  - *training algorithm*
  - *learner*
  - *algorithm*
  - *agent*

The statistical model

Assumption

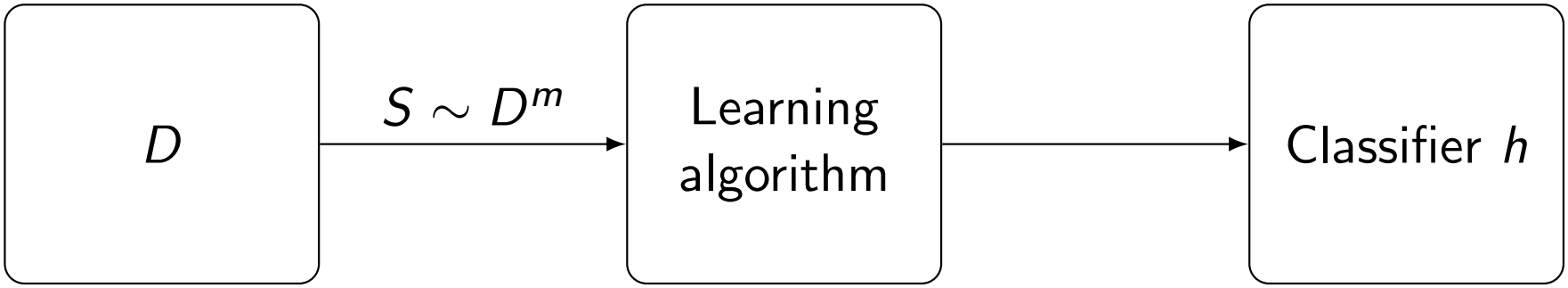
- *There exists a probability distribution D over $\mathcal{X} \times \mathcal{Y}$.*
- *Learning algorithm receives as an input a i.i.d. sample*

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

generated from D

- A random draw from D is denoted $(X, Y) \sim D$ or $(x, y) \sim D$
- Usually, random variables are denoted by capital letters. We will violate this often.
- i.i.d. = Independent and Identically Distributed
- S is sample drawn at random from a product distribution D^m over $(\mathcal{X} \times \mathcal{Y})^m$. This fact is denoted as $S \sim D^m$.

The statistical model



Notation and definitions #3

- A classifier h makes a *mistake* on a labeled example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ if and only if

$$h(x) \neq y$$

- *Sample error* of h on labeled sample $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$ is

$$L_S(h) = \frac{|\{i \ : \ 1 \leq i \leq m, \ h(x_i) = y_i\}|}{m} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\left[h(x_i) \neq y_i\right].$$

  Many different names depending on the context:
  - sample error
  - training error
  - test error
  - validation error
- Note that $L_S(h)$ is number between 0 and 1.

Notation and definitions #4

- *Generalization error* of h on a probability distribution over D over $\mathcal{X} \times \mathcal{Y}$

$$L_D(h) = \Pr_{(x,y)\sim D}[h(x) \neq y] = \mathbf{E}[\mathbf{1}[h(x) \neq y]] .$$

  Many different names:
  - *generalization error*
  - *true error*
  - *error*
  - *risk*
  - INCORRECT and WRONGLY used names:
    - test error
    - validation error
- Note that $L_D(h)$ is number between 0 and 1.

Notation and definitions #5

- The function $\ell_{0-1} : \mathcal{Y}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$,

$$\ell_{0-1}(h, (x, y)) = \mathbf{1}\left[h(x) \neq y\right]$$

  is called *zero-one loss*.
- This way,

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell_{0-1}(h, (x_i, y_i)) \,,$$
$$L_D(h) = \mathbf{E}_{(x,y)\sim D}\left[\ell_{0-1}(h, (x, y))\right] \,.$$

- Often other loss functions are used for various reasons.
- We will use *zero-one loss* for now.

The main goal of statistical learning

Main Goal

Produce a classifier $h : \mathcal{X} \to \mathcal{Y}$ with generalization error $L_D(h)$ that is as small as possible.

- In most scenarios, distribution D is unknown.
- In many scenarios, the distribution D is even unknowable. That is, D will never be known. *For example, number of black-and-white images with dimensions* $256x256$ *is* 2^{65536} *which is more than the number of atoms in the visible universe.*
- We will have access only to a labeled i.i.d. sample $S \sim D^m$.
- Is there a hope to find a classifier h with a small generalization error?

Bayes optimal classifier

- What is the classifier with the smallest generalization error?
- The classifier is called *Bayes optimal classifier*.
- For mathematical simplicity, assume the domain X is finite or countable.

Bayes optimal classifier

- Suppose $\mathcal{Y} = \{+1, -1\}$ and suppose $(X, Y) \sim D$
- Construct the conditional label distribution $f : \mathcal{X} \to [0, 1]$,

$$f(x) = \Pr[Y = 1 \mid X = x] \qquad \text{for any } x \in \mathcal{X}.$$

- *Bayes optimal classifier* is $h^* : \mathcal{X} \to \mathcal{Y}$ where

$$h^*(x) = \begin{cases} +1 & \text{if } f(x) \geq 1/2, \\ -1 & \text{if } f(x) < 1/2. \end{cases}$$

- Construct the conditional expectation $g : \mathcal{X} \to [-1, 1]$

$$g(x) = \mathbf{E}[Y \mid X = x] \qquad \text{for any } x \in \mathcal{X}.$$

- Equivalently, *Bayes optimal classifier* is $h^* : \mathcal{X} \to \mathcal{Y}$ where

$$h^*(x) = \begin{cases} +1 & \text{if } g(x) \geq 0, \\ -1 & \text{if } g(x) < 0. \end{cases}$$

Bayes optimal classifier

- Suppose $\mathcal{Y} = \{0, 1\}$ and suppose $(X, Y) \sim D$
- *Bayes optimal classifier* is $h^* : \mathcal{X} \to \mathcal{Y}$ where

$$h^*(x) = \begin{cases} 1 & \text{if } \Pr[Y = 1 \mid X = x] \geq 1/2, \\ 0 & \text{if } \Pr[Y = 1 \mid X = x] < 1/2. \end{cases}$$

- Equivalently, *Bayes optimal classifier* is $h^* : \mathcal{X} \to \mathcal{Y}$ where

$$h^*(x) = \begin{cases} 1 & \text{if } \mathbf{E}[Y \mid X = x] \geq 1/2, \\ 0 & \text{if } \mathbf{E}[Y \mid X = x] < 1/2. \end{cases}$$

- Note that $\Pr[Y = 1 \mid X = x] = \mathbf{E}[Y \mid X = x]$.

Bayes optimal classifier

Theorem (Optimality of Bayes optimal classifier)

Let $\mathcal{X}$ be a non-empty set. Let $\mathcal{Y} = \{0, 1\}$. Let D be a distribution over $\mathcal{X} \times \mathcal{Y}$. Let $h^* : \mathcal{X} \to \mathcal{Y}$ be the Bayes optimal classifier for D. Then, for any classifier $h : \mathcal{X} \to \mathcal{Y}$,

$$L_D(h^*) \leq L_D(h) \,.$$

Bayes optimal classifier

Proof (part 1)

Let $(X, Y) \sim D$. Then,

$$L_D(h^*) = \mathbf{E}[\ell_{0-1}(h^*, (X, Y))] = \mathbf{E}\left[\mathbf{E}[\ell_{0-1}(h^*, (X, Y)) \mid X]\right]$$
$$L_D(h) = \mathbf{E}[\ell_{0-1}(h, (X, Y))] = \mathbf{E}\left[\mathbf{E}[\ell_{0-1}(h, (X, Y)) \mid X]\right]$$

To prove $L_D(h^*) \leq L_D(h)$ it sufices to prove (why?)

$$\mathbf{E}[\ell_{0-1}(h^*, (X, Y)) \mid X] \leq \mathbf{E}[\ell_{0-1}(h, (X, Y)) \mid X] \,.$$

Equivalently,

$$\Pr[h^*(X) \neq Y \mid X] \leq \Pr[h(X) \neq Y \mid X] \,.$$

Bayes optimal classifier

Proof (part 2)

The inequality

$$\Pr[h^*(X) \neq Y \mid X] \leq \Pr[h(X) \neq Y \mid X]$$

is equivalent to

$$\begin{aligned} &\Pr[h^*(X) = 1, Y = 0 \mid X] + \Pr[h^*(X) = 0, Y = 1 \mid X] \\ &\quad \leq \Pr[h(X) = 1, Y = 0 \mid X] + \Pr[h(X) = 0, Y = 1 \mid X] \,. \end{aligned} \tag{1}$$

If $h^*(X) = h(X)$, the inequality (1) holds with equality.

If $h^*(X) \neq h(X)$, we have two cases.

Bayes optimal classifier

Proof (part 3)

Case 1: $h^*(X) = 1, h(X) = 0$. Inequality (1) simplifies to

$$\begin{aligned} &\Pr[h^*(X) = 1, Y = 0 \mid X] + \cancel{\Pr[h^*(X) = 0, Y = 1 \mid X]} \\ &\quad \leq \cancel{\Pr[h(X) = 1, Y = 0 \mid X]} + \Pr[h(X) = 0, Y = 1 \mid X] \,. \end{aligned} \tag{1}$$

Thus, we need to prove

$$\Pr[Y = 0 \mid X] \leq \Pr[Y = 1 \mid X] \,,$$

This follows from the definition of h^* and that $h^*(X) = 1$.

Bayes optimal classifier

Proof (part 4)

Case 2: $h^*(X) = 0, h(X) = 1$. Inequality (1) simplifies to

$$\begin{aligned} &\cancel{\Pr[h^*(X) = 1, Y = 0 \mid X]} + \Pr[h^*(X) = 0, Y = 1 \mid X] \\ &\quad \leq \Pr[h(X) = 1, Y = 0 \mid X] + \cancel{\Pr[h(X) = 0, Y = 1 \mid X]} \,. \end{aligned} \tag{1}$$

Thus, we need to prove

$$\Pr[Y = 1 \mid X] \leq \Pr[Y = 0 \mid X] \,,$$

This follows from the definition of h^* and that $h^*(X) = 1$. ■