Statistical Machine Learning, Part I

Statistical Learning Theory

mcuturi@i.kyoto-u.ac.jp

Previous Lecture : Classification

- Classification: mapping objects onto $\mathcal{S}$ where $|\mathcal{S}| < \infty$.
- Binary classification: answers to **yes/no** questions
- Linear classification algorithms: *split* the **yes/no** zones with a **hyperplane**

$$\text{Yes} = \{\mathbf{c}^T x + \boldsymbol{b} \geq 0\}\ ,\ \text{No} = \{\mathbf{c}^T x + \boldsymbol{b} < 0\}$$

- How to select $\mathbf{c}, \boldsymbol{b}$ given a dataset?
  - **Linear Discriminant Analysis** (multivariate Gaussians)
  - **Logistic Regression** (classification from a linear regression viewpoint)
  - **Perceptron rule** (iterative, random update rule)
  - brief introduction to **Support Vector Machine** (optimal margin classifier)

Today

- Usual steps when using ML algorithms
  - Define problem (*classification? regression? multi-class?*)
  - Gather data
  - Choose representation for data to build a database
  - Choose method/algorithm based on training set
  - Choose/estimate parameters
  - Run algorithm on new points, collect results

Today

- Usual steps when using ML algorithms
  - Define problem (*classification? regression? multi-class?*)
  - Gather data
  - Choose representation for data to build a database
  - **Choose method/algorithm**
  - **Choose/estimate parameters based on training set**
  - Run algorithm on new points, collect results
  - ... ***did I overfit?***

Probabilistic Framework

General Framework

- Couples of observations, $(\mathbf{x}, y)$ appear in nature.
- These observations are
$$\mathbf{x} \in \mathbb{R}^d, \quad y \in \mathcal{S}$$
- $\mathcal{S} \subset \mathbb{R}$, that is $\mathcal{S}$ could be $\mathbb{R}$, $\mathbb{R}_+$, $\{1, 2, 3, \dots, L\}$, $\{0, 1\}$
- Sometimes only $\mathbf{x}$ is visible. We want to guess the most likely y for that $\mathbf{x}$.

- **Example 1** $\mathbf{x}$: Height $\in \mathbb{R}$, y: Gender $\in \{M, F\}$

X is 164cm tall, is X a male or a female?

- **Example 2** $\mathbf{x}$: Height $\in \mathbb{R}$, y: Weight $\in \mathbb{R}$.

X is 164cm tall, how many kilos does X weight?

Estimating the relationship between x and y

- To provide a guess $\Leftrightarrow$ estimate a function $f : \mathbb{R}^d \to \mathcal{S}$ such that

$$f(\mathbf{x}) \approx y.$$

Estimating the relationship between x and y

- To provide a guess $\Leftrightarrow$ estimate a function $f : \mathbb{R}^d \rightarrow \mathcal{S}$ such that

$$f(\mathbf{x}) \approx y.$$

- Ideally, $f(\mathbf{x}) \approx y$ should apply **both** to
  - couples $(\mathbf{x}, y)$ we **have observed** in the training set
  - couples $(\mathbf{x}, y)$ we **will observe**... (guess y from $\mathbf{x}$)

Probabilistic Framework

- We **assume** that **each** observation $(\mathbf{x}, y)$ arises as an
  - **independent**,
  - **identically distributed**,

  random sample from the **same** probability law.

Probabilistic Framework

- We **assume** that **each** observation $(\mathbf{x}, y)$ arises as an
  - **independent**,
  - **identically distributed**,

  random sample from the **same** probability law.

- This probability P on $\mathbb{R}^d \times \mathcal{S}$ has a density,

$$p(X = \mathbf{x}, Y = y).$$

Probabilistic Framework

- We **assume** that **each** observation $(\mathbf{x}, y)$ arises as an
  - **independent**,
  - **identically distributed**,

  random sample from the **same** probability law.

- This probability P on $\mathbb{R}^d \times \mathcal{S}$ has a density,

$$p(X = \mathbf{x}, Y = y).$$

- This also provides us with the **marginal** probabilities for $\mathbf{x}$ and y:

$$p(Y = y) = \int_{\mathbb{R}^d} p(X = \mathbf{x}, Y = y) d\mathbf{x}$$

$$p(X = \mathbf{x}) = \int_{\mathcal{S}} p(X = \mathbf{x}, Y = y) dy$$

Probabilistic Framework

- Assuming that p **exists** is fundamental in statistical learning theory.

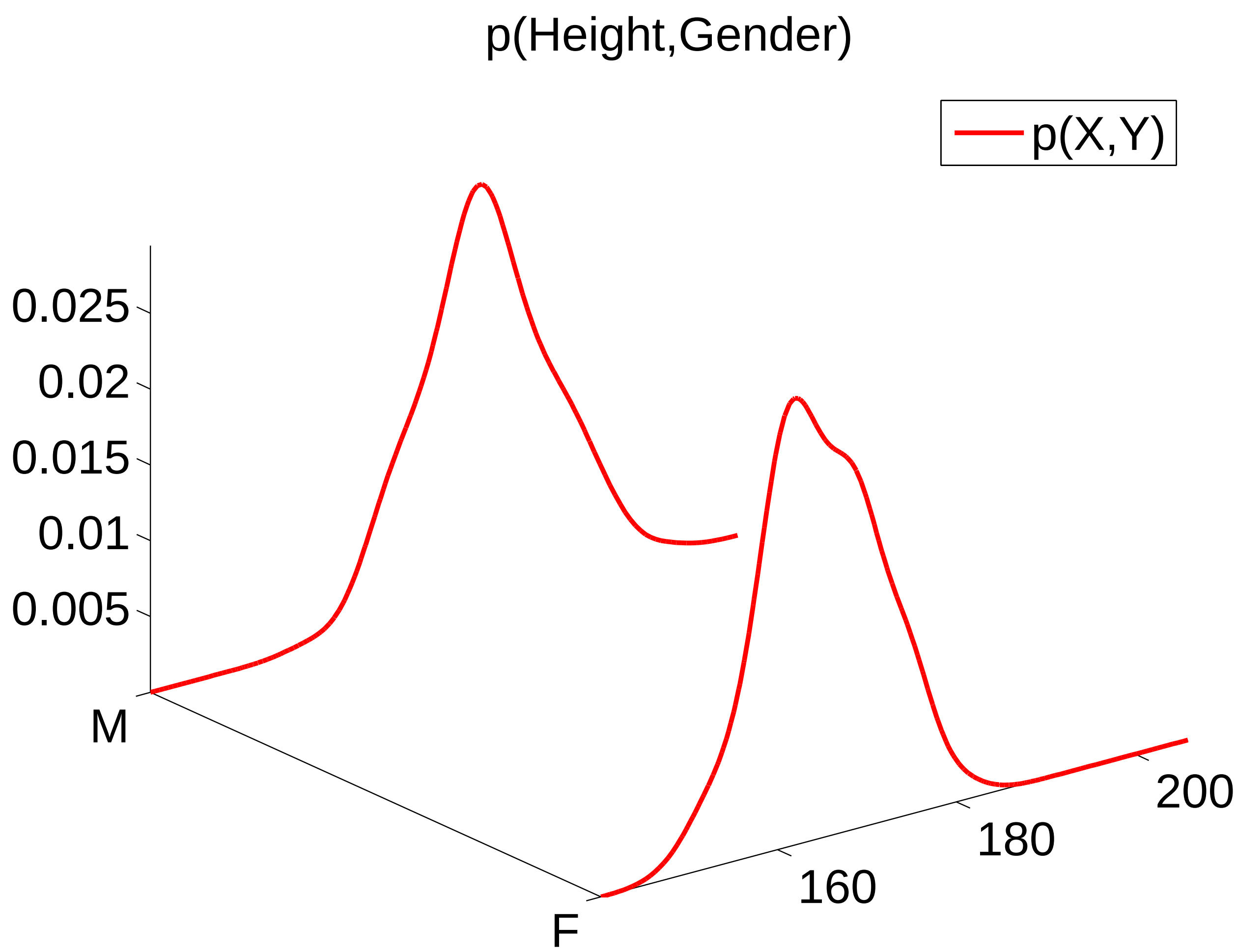
$$p(X = \mathbf{x}, Y = y).$$

- What happens to learning problems if **we know** p?..
(**in practice**, this will **never** happen, we never know p).

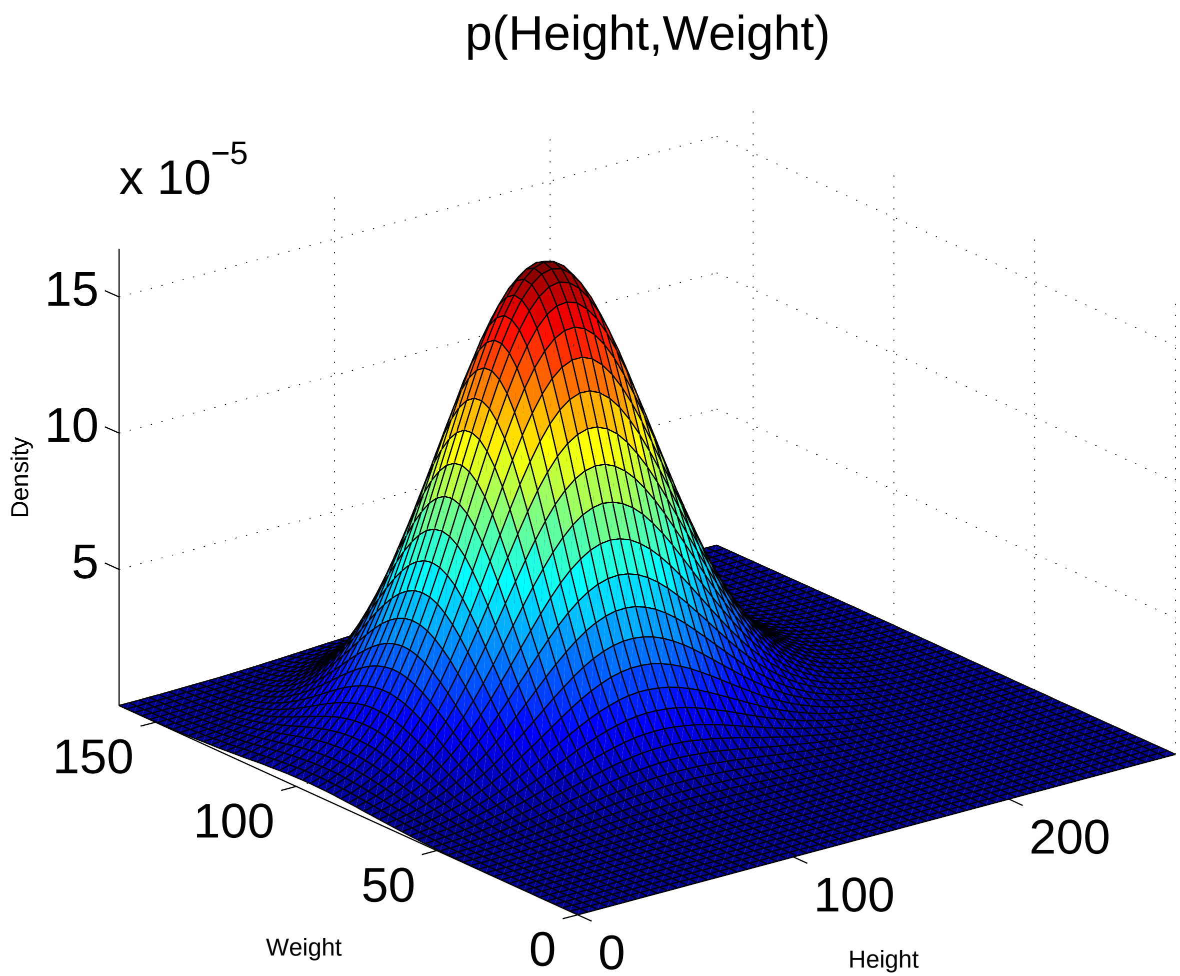
- If we know p, learning problems become **trivial**.

(≈ running a marathon on a motorbike)

Example 1: $\mathcal{S} = \{M, F\}$, Height vs Gender



Example 2: $\mathcal{S} = \mathbb{R}^+$, Height vs Weight



Probabilistic Framework

Conditional probability (or density)

$$p(A, B) = p(A|B)p(B)$$

- Suppose:

$$p(X = 184\text{cm}, y = M) = 0.015$$
$$p(y = M) = 0.5$$

What is $p(X = 184\text{cm} \mid y = M)$?

- 1. 0.15
- 2. 0.03
- 3. 0.5
- 4. 0.0075
- 5. 0.2

Probabilistic Framework

Bayes Rule

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

- Suppose:

$$p(X = 184\text{cm} \mid y = M) = 0.03$$
$$p(y = M) = 0.5$$
$$p(X = 184) = 0.02$$

  What is $p(y = M|X = 184)$?

  - 1. 0.6
  - 2. 0.04
  - 3. 0.75
  - 4. 0.8
  - 5. 0.2

Loss, Risk and Bayes Decision

Building Blocks: Loss (1)

- A loss is a function $\mathcal{S} \times \mathbb{R} \to \mathbb{R}_+$ designed to **quantify** mistakes,

how good is the prediction $f(\mathbf{x})$ given that the true answer is y? $\Updownarrow$ How small is $l(y, f(\mathbf{x}))$?

Examples

- $\mathcal{S} = \{0, 1\}$
  - 0/1 loss: $l(a, b) = \delta_{a \neq b} = \begin{cases} 1 \text{ if } a \neq b \\ 0 \text{ if } a = b \end{cases}$
- $\mathcal{S} = \mathbb{R}$
  - Squared euclidian distance $l(a, b) = (a - b)^2$
  - norm $l(a, b) = \|a - b\|_q$, $0 \leq q \leq \infty$

Building Blocks: Risk (2)

- The **Risk** of a predictor $\boldsymbol{f}$ with respect to **loss** l is

$$R(f) = \mathbb{E}_{\boldsymbol{p}}[l(Y, \boldsymbol{f}(X))] = \int_{\mathbb{R}^d \times \mathcal{S}} \boldsymbol{l}(y, \boldsymbol{f}(\mathbf{x}))\, \boldsymbol{p}(\mathbf{x}, \boldsymbol{y}) d\mathbf{x} dy$$

- Risk $=$ average loss of $\boldsymbol{f}$ on **all possible couples** $(\mathbf{x}, y)$,

weighted by the probability density.

Risk($\boldsymbol{f}$) measures the performance of f w.r.t. $\boldsymbol{l}$ and $\boldsymbol{p}$.

- Remark: a function $\boldsymbol{f}$ with **low risk** can make **very big mistakes** for some $\mathbf{x}$ as long as the **probability** $p(\mathbf{x})$ of $\mathbf{x}$ is **small**.

A lower bound on the Risk? Bayes Risk

- Since $l \geq 0$, $R(\boldsymbol{f}) \geq 0$.
- Consider all possible functions $\mathbb{R}^d \to \mathcal{S}$, usually written $(\mathbb{R}^d)^{\mathcal{S}}$.
- The **Bayes** risk is the quantity

$$R^* = \inf_{\boldsymbol{f} \in (\mathbb{R}^d)^{\mathcal{S}}} R(\boldsymbol{f}) = \inf_{\boldsymbol{f} \in (\mathbb{R}^d)^{\mathcal{S}}} \mathbb{E}_p[l(Y, \boldsymbol{f}(X))]$$

- Ideal classifier would have Bayes risk.

Bayes Classifier : $\mathcal{S} = \{0, 1\}$, l is the 0/1 loss.

Let's write: $\eta(\mathbf{x}) = p(Y = 1 | X = \mathbf{x})$.

- Define the following rule:

$$g_B(\mathbf{x}) = \begin{cases} 1, & \text{if } \eta(\mathbf{x}) \geq \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

where

The **Bayes classifier** achieves the **Bayes Risk**.

Theorem 1. $R(g_B) = R^*$.

Bayes Classifier : $\mathcal{S} = \{0, 1\}$, l is the 0/1 loss.

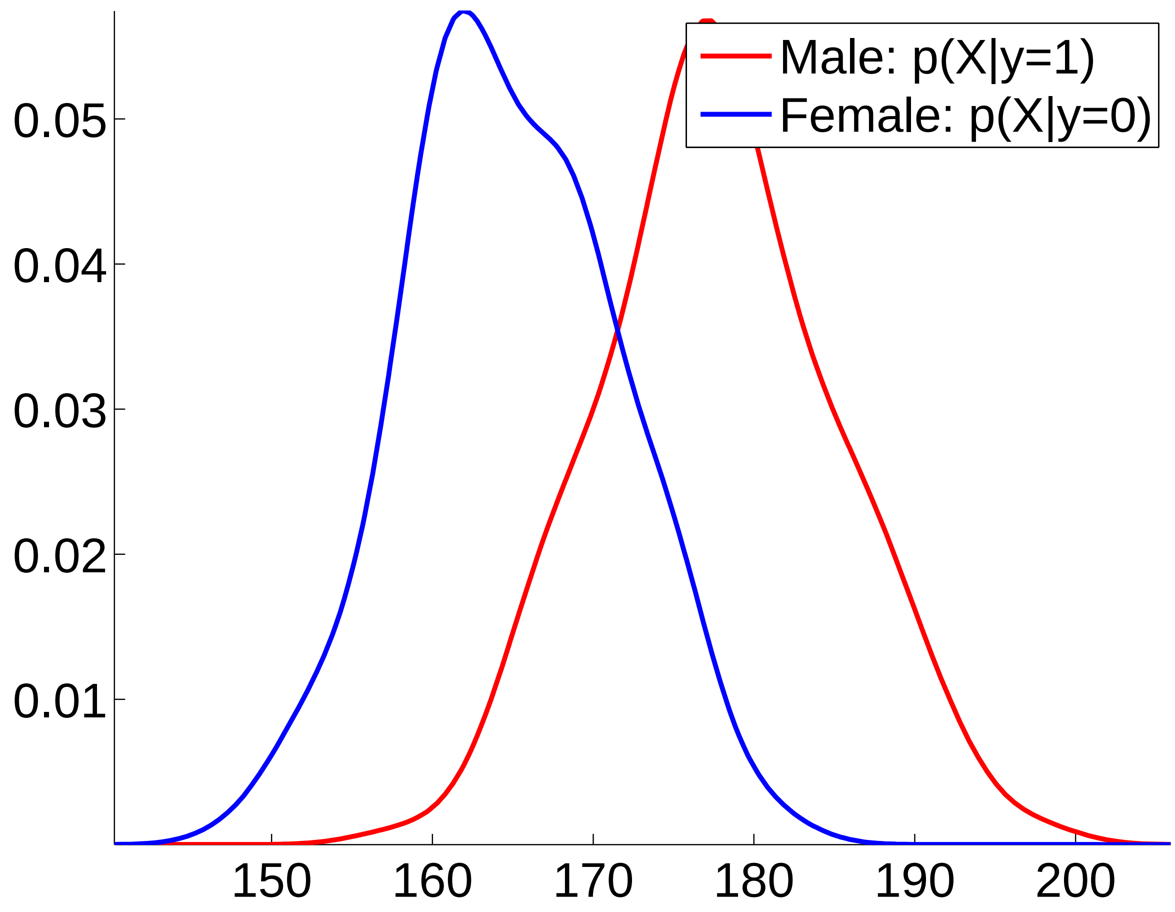
- Chain rule of conditional probability $p(A, B) = p(B)p(A|B)$
- Bayes rule

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

- A simple way to compute η:

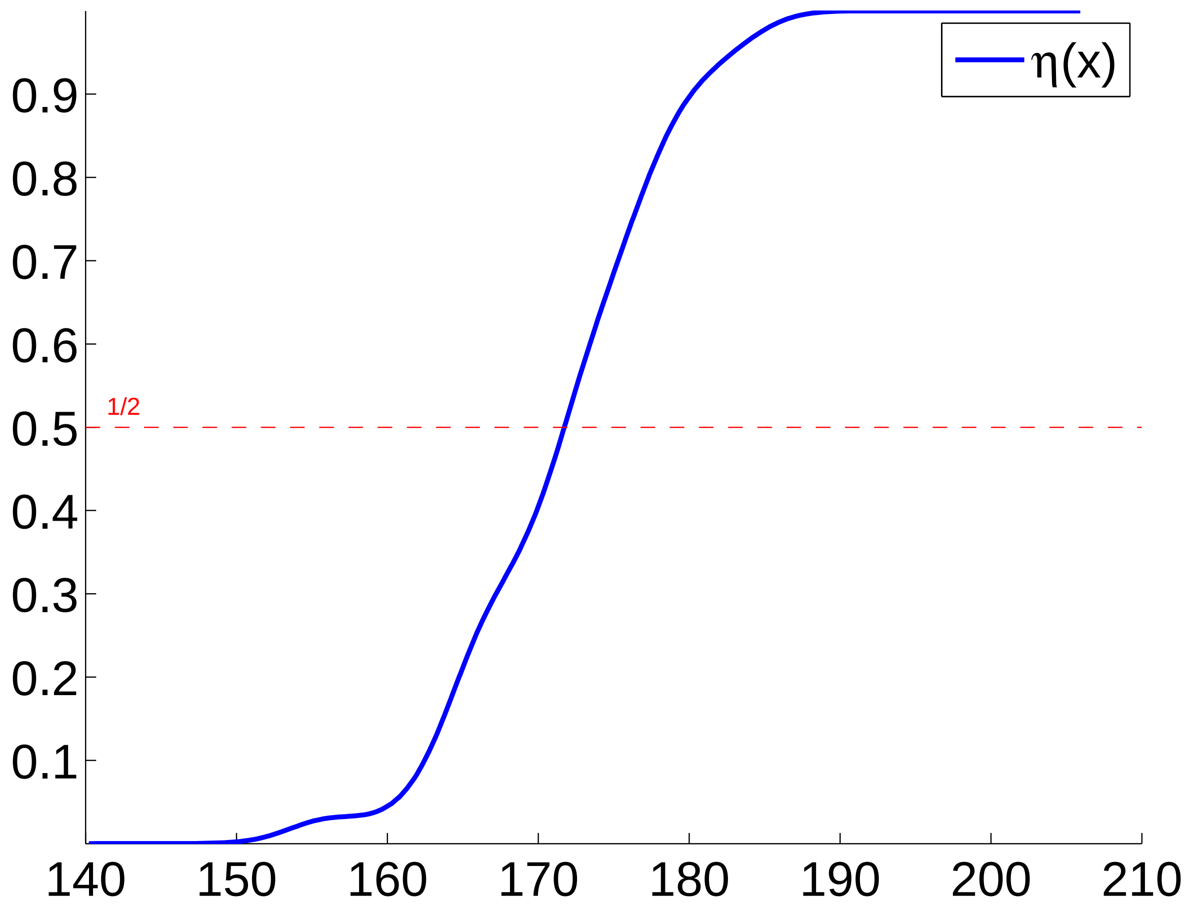
$$\begin{aligned}\eta(\mathbf{x}) = p(Y = 1|X = \mathbf{x}) &= \frac{p(Y = 1, X = \mathbf{x})}{p(X = \mathbf{x})} \\ &= \frac{p(X = \mathbf{x}|Y = 1)p(Y = 1)}{p(X = \mathbf{x})} \\ &= \frac{p(X = \mathbf{x}|Y = 1)p(Y = 1)}{p(X = \mathbf{x}|Y = 1)p(Y = 1) + p(X = \mathbf{x}|Y = 0)p(Y = 0)}.\end{aligned}$$

Bayes Classifier : $\mathcal{S} = \{0, 1\}$, l is the 0/1 loss.



in addition, $p(Y = 1) = 0.4871$. As a consequence
$p(Y = 0) = 1 - 0.4871 = 0.5129$

Bayes Classifier : $\mathcal{S} = \{0, 1\}$, l is the 0/1 loss.



Bayes Estimator : $\mathcal{S} = \mathbb{R}$, l is the 2-norm

- Consider the following rule:

$$g_B(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}] = \int_{\mathbb{R}} y\, p(Y = y|X = \mathbf{x})dy$$

Here again, the **Bayes estimator** achieves the **Bayes Risk**.

Theorem 2. $R(g_B) = R^*$.

Bayes Estimator : $\mathcal{S} = \mathbb{R}$, l is the 2-norm

- Using Bayes rule again,

$$
\begin{aligned}
f^\star(\mathbf{x}) = \mathbb{E}[Y|X=\mathbf{x}] &= \int_{\mathbb{R}} \boldsymbol{y}\, p(Y=y|X=\mathbf{x})dy \\
&= \int_{\mathbb{R}} \boldsymbol{y}\, \frac{p(X=\mathbf{x}|Y=y)p(Y=y)}{p(X=\mathbf{x})}dy \\
&= \int_{\mathbb{R}} \boldsymbol{y}\, \frac{p(X=\mathbf{x}|Y=y)p(Y=y)}{\int_{\mathbb{R}} p(X=\mathbf{x}|Y=u)p(Y=u)du}dy \\
&= \frac{\int_{\mathbb{R}} \boldsymbol{y}\, p(X=\mathbf{x}|Y=y)p(Y=y)dy}{\int_{\mathbb{R}} p(X=\mathbf{x}|Y=y)p(Y=y)dy}
\end{aligned}
$$

In practice: No p, Only Finite Samples

What can we do?

- If we know the probability p, Bayes estimator would be impossible to beat.
- In practice, the only thing we can use is a training set,

$$\{(\mathbf{x}_i, y_i)\}_{i=1,\cdots,n}.$$

- For instance, a list of Heights, gender

163.0000	F
170.0000	F
175.3000	M
184.0000	M
175.0000	M
172.5000	F
153.5000	F
164.0000	M
163.0000	M

Approximating Risk

- For any function f, we **cannot** compute its true risk $R(f)$,

$$R(f) = \mathbb{E}_p[l(Y, f(X))]$$

  because **we do not know** p

- Instead, we can consider the **empirical** Risk R_n^{emp}, defined as

$$R_n^{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(\mathbf{x}_i))$$

- The law of large numbers tells us that for any given f

$$R_n^{\text{emp}}(f) \to R(f).$$

Relying on the empirical risk

As sample size n grows, the empirical risk behaves like the *real* risk

- It *may* thus seem like a good idea to **minimize directly** the empirical risk.
- The intuition is that
  - since a function f such that $R(f)$ is low is desirable,
  - since $\boldsymbol{R_n^{\text{emp}}}(f)$ converges to $R(f)$ as $n \to \infty$,

why not look directly for any function f such that $\boldsymbol{R_n^{\text{emp}}}(f)$ is low?

- Typically, in the context of classification with 0/1 loss, find a function such that

$$\boldsymbol{R_n^{\text{emp}}}(f) = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i \neq f(\mathbf{x}_i)}$$

...is low.

A flawed intuition

- However, focusing **only** on R_n^{emp} is not viable.
- Many ways this can go wrong...

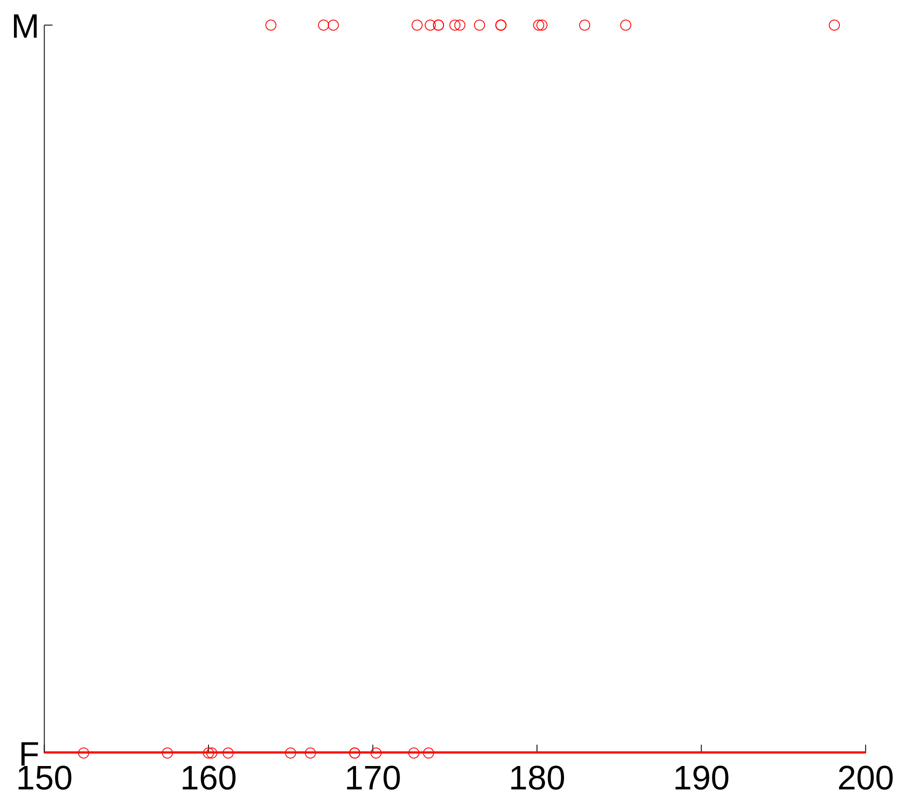
A flawed intuition

- Consider the function defined as

$$h(\mathbf{x}) = \begin{cases} y_1, & \text{if } \mathbf{x} = \mathbf{x}_1, \\ y_2, & \text{if } \mathbf{x} = \mathbf{x}_2, \\ \vdots & \\ y_n, & \text{if } \mathbf{x} = \mathbf{x}_n, \\ 0 \text{ otherwise..} & \end{cases}$$

- Since, $\boldsymbol{R}_{\boldsymbol{n}}^{\mathbf{emp}}(h) = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i \neq h(\mathbf{x}_i)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i \neq y_i} = 0$, h minimizes $\boldsymbol{R}_{\boldsymbol{n}}^{\mathbf{emp}}$.
- However, h **always** answers **0**, except for a few points.
- In practice, we can expect $R(h)$ to be much higher, equal to $P(Y = 1)$ in fact.

Here is what this function would predict on the Height/Gender Problem



Overfitting is probably the **most frequent mistake** made by ML practitioners.

Ideas to Avoid Overfitting

- Our criterion $\boldsymbol{R}_{\boldsymbol{n}}^{\textbf{emp}}(g)$ only considers a **finite** set of points.
- A function g defined on $\mathbb{R}^d$ is defined on an **infinite** set of points.

A few approaches to control overfitting

- **Restrict the set of candidates**

$$\min_{g \in \boldsymbol{\mathcal{G}}} \boldsymbol{R}_{\boldsymbol{n}}^{\textbf{emp}}(g).$$

- **Penalize "undesirable" functions**

$$\min_{g \in \boldsymbol{\mathcal{G}}} \boldsymbol{R}_{\boldsymbol{n}}^{\textbf{emp}}(g) + \lambda \|\boldsymbol{g}\|^{\mathbf{2}}$$

Are there theoretical tools which justify such approaches?

Bounds

Flow of a learning process in Machine Learning

- Assumption 1. existence of a probability density p for (X, Y).
- Assumption 2. points are observed i.i.d. following this probability density.

Flow of a learning process in Machine Learning

- Assumption 1. existence of a probability density p for (X, Y).
- Assumption 2. points are observed i.i.d. following this probability density.

Roadmap

- Get a random training sample $\{(\mathbf{x}_j, y_j)\}_{i=1,\cdots,n}$ (*training set*)
- Choose a class of functions $\mathcal{G}$ (*method* or *model*)
- Choose g_n in $\mathcal{G}$ such that $\boldsymbol{R_n^{\mathrm{emp}}}(g_n)$ is **low** (*estimation algorithm*)

Flow of a learning process in Machine Learning

- Assumption 1. existence of a probability density p for (X, Y).
- Assumption 2. points are observed i.i.d. following this probability density.

Roadmap

- Get a random training sample $\{(\mathbf{x}_j, y_j)\}_{i=1,\cdots,n}$ (*training set*)
- Choose a class of functions $\mathcal{G}$ (*method* or *model*)
- Choose g_n in $\mathcal{G}$ such that $\boldsymbol{R_n^{\text{emp}}}(g_n)$ is **low** (*estimation algorithm*)

Next... use g_n in practice

Flow of a learning process in Machine Learning

Yet, you may want to have a partial answer to these questions

- How good would be $\boldsymbol{g_B}$ if we knew the real probability p?
- what about $\boldsymbol{R(g_n)}$?
- What's the gap between them, $\boldsymbol{R(g_n)} - \boldsymbol{R}(\boldsymbol{g_B})$?
- Is the *estimation* algorithm reliable? how big is $\boldsymbol{R^{\mathrm{emp}}(g_n)} - \inf_{g\in\mathcal{G}} \boldsymbol{R_n^{\mathrm{emp}}}(g)$?
- how big is $\boldsymbol{R_n^{\mathrm{emp}}}(g_n) - \inf_{g\in\mathcal{G}} \boldsymbol{R}(g)$?

Excess Risk

- In the general case $g_B \notin \mathcal{G}$.
- Hence, by introducing $g^\star$ as a function achieving the lowest risk in $\mathcal{G}$,

$$R(g^\star) = \inf_{g \in \mathcal{G}} R(g),$$

we decompose

$$R(g_n) - R(g_B) = [R(g_n) - R(g^\star)] + [R(g^\star) - R(g_B)]$$

Excess Risk

- In the general case $g_B \notin \mathcal{G}$.
- Hence, by introducing $g^\star$ as a function achieving the lowest risk in $\mathcal{G}$,

$$R(g^\star) = \inf_{g \in \mathcal{F}} R(g),$$

we decompose

$$R(g_n) - R(g_B) = \underbrace{\left[\boldsymbol{R(g_n) - R(g^\star)}\right]}_{\textbf{Estimation Error}} + \underbrace{\left[\boldsymbol{R(g^\star) - R(g_B)}\right]}_{\textbf{Approximation Error}}$$

- Estimation error is **random**, Approximation error is **fixed**.
- In the following we focus on the estimation error.

Types of Bounds

Error Bounds

$$R(g_n) \leq \boldsymbol{R_n^{\mathrm{emp}}}(g_n) + C(n, \mathcal{G}).$$

Types of Bounds

Error Bounds

$$R(g_n) \leq \boldsymbol{R_n^{\text{emp}}}(g_n) + C(n, \mathcal{G}).$$

Error Bounds Relative to Best in Class

$$R(g_n) \leq R(g^\star) + C(n, \mathcal{G}).$$

Types of Bounds

Error Bounds

$$R(g_n) \leq \boldsymbol{R_n^{\mathrm{emp}}}(g_n) + C(n, \mathcal{G}).$$

Error Bounds Relative to Best in Class

$$R(g_n) \leq R(g^\star) + C(n, \mathcal{G}).$$

Error Bounds Relative to the Bayes Risk

$$R(g_n) \leq R(g_B) + C(n, \mathcal{G}).$$

Error Bounds / Generalization Bounds

$$R(g_n) - \boldsymbol{R_n^{\text{emp}}}(g_n)$$

What is Overfitting?

- Overfitting is the idea that,
  - given n training points sampled randomly,
  - given a function g_n estimated from these points,
  - we may have...

$$R(g_n) \gg \boldsymbol{R_n^{\mathrm{emp}}}(g_n).$$

What is Overfitting?

- Overfitting is the idea that,
  - given n training points sampled randomly,
  - given a function g_n estimated from these points,
  - we may have...

$$R(g_n) \gg \boldsymbol{R_n^{\text{emp}}}(g_n).$$

- Question of interest:

$$P[R(g_n) - \boldsymbol{R_n^{\text{emp}}}(g_n) > \varepsilon] = ?$$

- From now on, we consider the **classification** case, namely $\mathcal{G} : \mathbb{R}^d \to \{0, 1\}$.

Alleviating Notations

- More convenient to see a couple $(\mathbf{x}, y)$ as a realization of Z, namely

$$\mathbf{z}_i = (\mathbf{x}_i, y_i), Z = (X, Y).$$

- We define the *loss class*

$$\mathcal{F} = \{f : \mathbf{z} = (\mathbf{x}, y) \rightarrow \delta_{g(\mathbf{x}) \neq y},\ g \in \mathcal{G}\},$$

- with the additional notations

$$Pf = \mathbb{E}[f(X, Y)], P_n f = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i, y_i),$$

where we recover

$$P_n f = \boldsymbol{R_n^{\mathrm{emp}}}(g), \quad Pf = R(g)]$$

Empirical Processes

For each $f \in \mathcal{F}$, $P_n f$ is a random variable which depends on n realizations of Z.

- If we consider **all** possible functions $f \in \mathcal{F}$, we obtain

The set of random variables $\{P_n f\}_{f \in \mathcal{F}}$ is called an Empirical measure indexed by $\mathcal{F}$.

- A branch of mathematics studies explicitly the convergence of $\{Pf - P_n f\}_{f \in \mathcal{F}}$,

This branch is known as Empirical process theory.

Hoeffding's Inequality

- Recall that for a given g and corresponding f,

$$R(g) - R^{\text{emp}}(g) = Pf - P_n f = \mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i),$$

  which is simply the difference between the **expectation** and the empirical average of $f(Z)$.

- The **strong** law of large numbers says that

$$P\left(\lim_{n\to\infty} \mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i) = 0\right) = 1.$$

Hoeffding's Inequality

- A more detailed result is

  Theorem 3 (Hoeffding). *Let* $Z_1, \cdots, Z_n$ *be* n *i.i.d random variables with* $f(Z) \in [a, b]$. *Then,* $\forall \varepsilon$,

$$P\left[|P_n f - Pf| > \varepsilon\right] \leq 2e^{-\frac{2n\varepsilon^2}{(b-a)^2}}.$$