Classification and Regression I

Krzysztof Dembczyński

Intelligent Decision Support Systems Laboratory (IDSS)
Poznań University of Technology, Poland

Software Development Technologies
Master studies, second semester
Academic year 2018/19 (winter course)

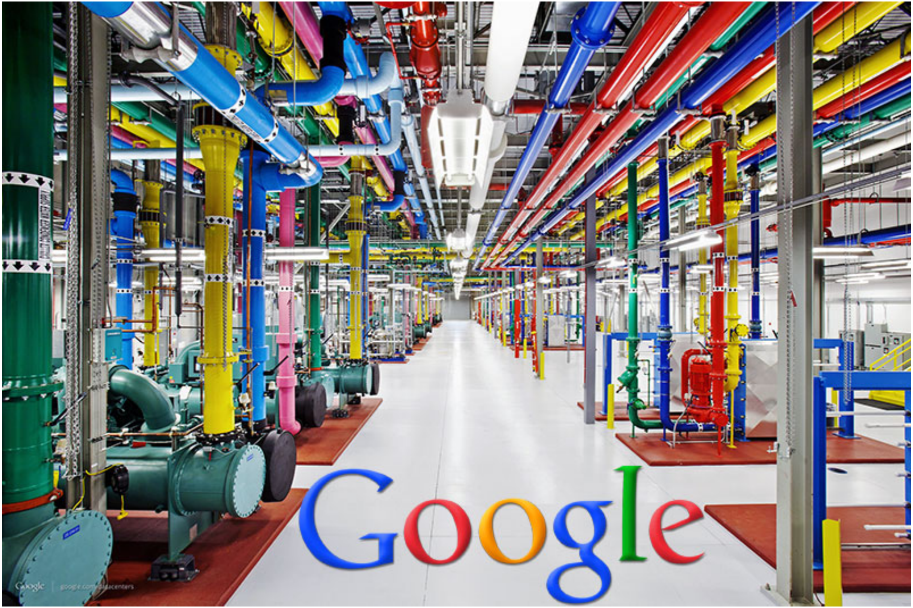
Review of the previous lectures

- Mining of massive datasets.

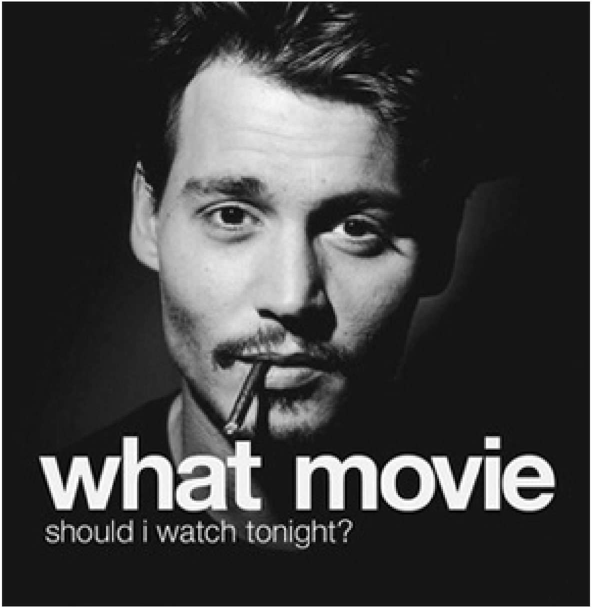
Outline

Outline

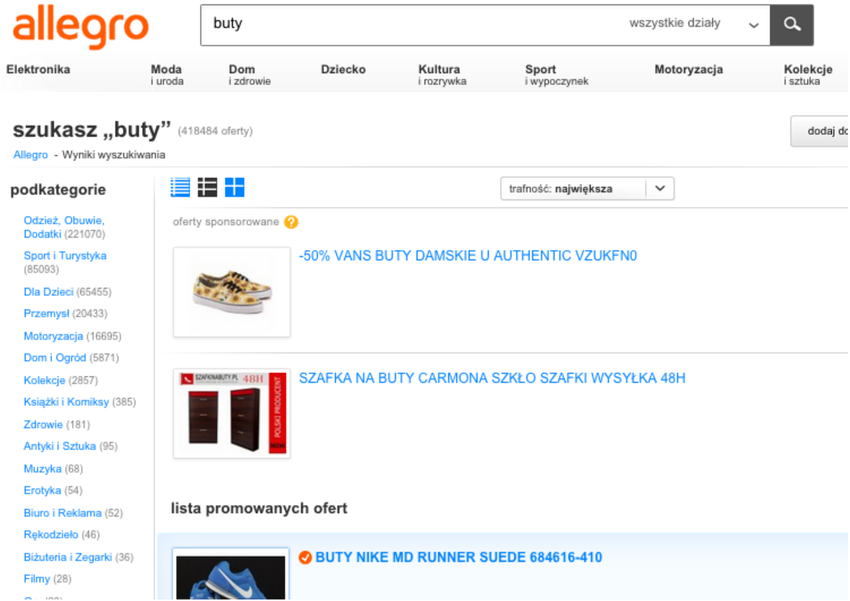
We live in the era of Big Data and Machine Learning.



Search engines: website ranking and personalization



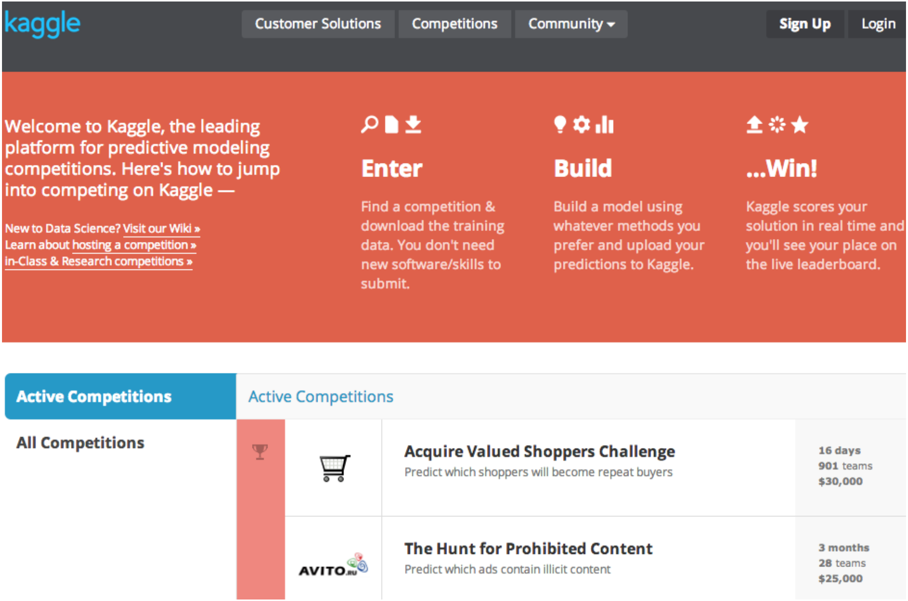
Recommender systems: movie, book, product recommendations



Online shopping/auctions

Autonomous vehicles

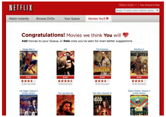
Spam filtering



A plenty of machine learning competitions

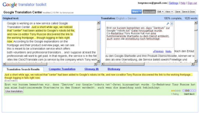
Machine learning is everywhere. . .

search engines

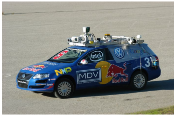


recommender systems

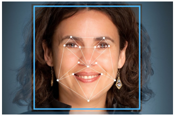
online advertising



translation

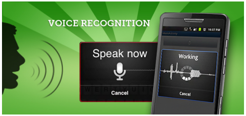


autonomous cars



face recognition

image recognition



voice recognition

fraud detection

healthcare

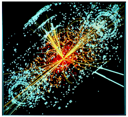
medical research

bioinformatics

neuroscience

climate science

astronomy



physics

Machine learning

- What is machine learning?

Machine learning

- What is machine learning?
  - Machine learning is the science of getting computers to act without being explicitly programmed. (Andrew Ng)

Machine learning

- What is machine learning?
  - Machine learning is the science of getting computers to act without being explicitly programmed. (Andrew Ng)
- Supervised learning
  - Learn a computer to predict an unknown response/value of a decision attribute for an object described by several features.

Machine learning

- What is machine learning?
  - Machine learning is the science of getting computers to act without being explicitly programmed. (Andrew Ng)
- Supervised learning
  - Learn a computer to predict an unknown response/value of a decision attribute for an object described by several features.
- Two main problems:
  - Classification: Prediction of categorical response,
  - Regression: Prediction of continuous response.

Machine learning

- What is machine learning?
  - Machine learning is the science of getting computers to act without being explicitly programmed. (Andrew Ng)
- Supervised learning
  - Learn a computer to predict an unknown response/value of a decision attribute for an object described by several features.
- Two main problems:
  - Classification: Prediction of categorical response,
  - Regression: Prediction of continuous response.
- Examples:
  - Spam filtering,
  - Handwriting recognition,
  - Text classification,
  - Stock prices,
  - etc.

Machine learning

- We know relatively much about solving simple learning problems such as binary classification:

Machine learning

- We know relatively much about solving simple learning problems such as binary classification:
  - Advanced theory,

Machine learning

- We know relatively much about solving simple learning problems such as binary classification:
  - Advanced theory,
  - Implemented fast algorithms,

Machine learning

- We know relatively much about solving simple learning problems such as binary classification:
  - Advanced theory,
  - Implemented fast algorithms,
  - Almost a mature technology.

Machine learning

- We know relatively much about solving simple learning problems such as binary classification:
  - Advanced theory,
  - Implemented fast algorithms,
  - Almost a mature technology.
- The main challenges are:

Machine learning

- We know relatively much about solving simple learning problems such as binary classification:
  - Advanced theory,
  - Implemented fast algorithms,
  - Almost a mature technology.
- The main challenges are:
  - Feature engineering,

Machine learning

- We know relatively much about solving simple learning problems such as binary classification:
  - Advanced theory,
  - Implemented fast algorithms,
  - Almost a mature technology.
- The main challenges are:
  - Feature engineering,
  - Supervision of examples,

Machine learning

- We know relatively much about solving simple learning problems such as binary classification:
  - Advanced theory,
  - Implemented fast algorithms,
  - Almost a mature technology.
- The main challenges are:
  - Feature engineering,
  - Supervision of examples,
  - New applications,

Machine learning

- We know relatively much about solving simple learning problems such as binary classification:
  - Advanced theory,
  - Implemented fast algorithms,
  - Almost a mature technology.
- The main challenges are:
  - Feature engineering,
  - Supervision of examples,
  - New applications,
  - Complex problems,

Machine learning

- We know relatively much about solving simple learning problems such as binary classification:
  - Advanced theory,
  - Implemented fast algorithms,
  - Almost a mature technology.
- The main challenges are:
  - Feature engineering,
  - Supervision of examples,
  - New applications,
  - Complex problems,
  - Large-scale problems.

Software

- Weka (http://www.cs.waikato.ac.nz/ml/weka/)
- R-project (http://www.r-project.org/),
- Octave (https://www.gnu.org/software/octave/),
- Julia (http://julialang.org/),
- Scikit-learn (http://scikit-learn.org/stable/)
- Matlab (http://www.mathworks.com/products/matlab/)
- H20 (http://0xdata.com/)
- GraphLab (http://dato.com/)
- MLLib (https://spark.apache.org/mllib/)
- Mahout (http://mahout.apache.org/)
- ...

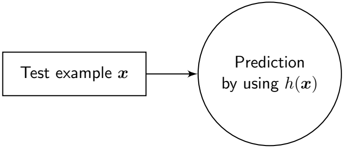
Outline

Supervised learning

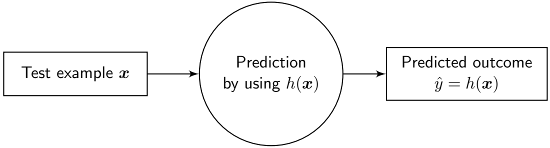
Supervised learning

Test example x

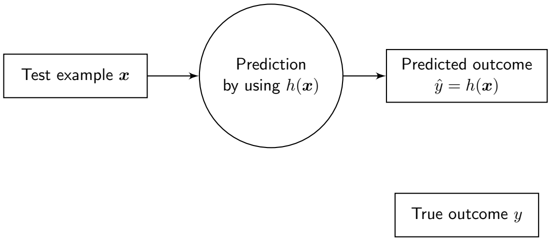
Supervised learning



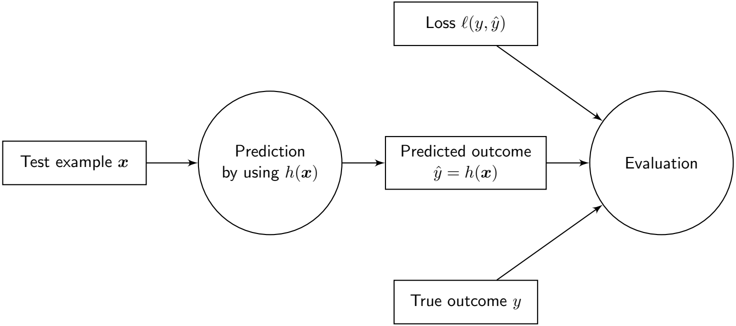
Supervised learning



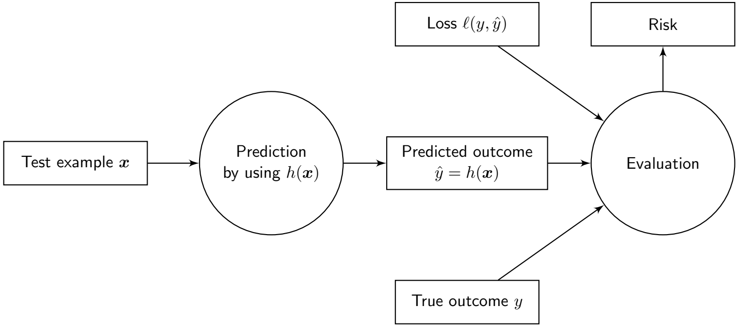
Supervised learning



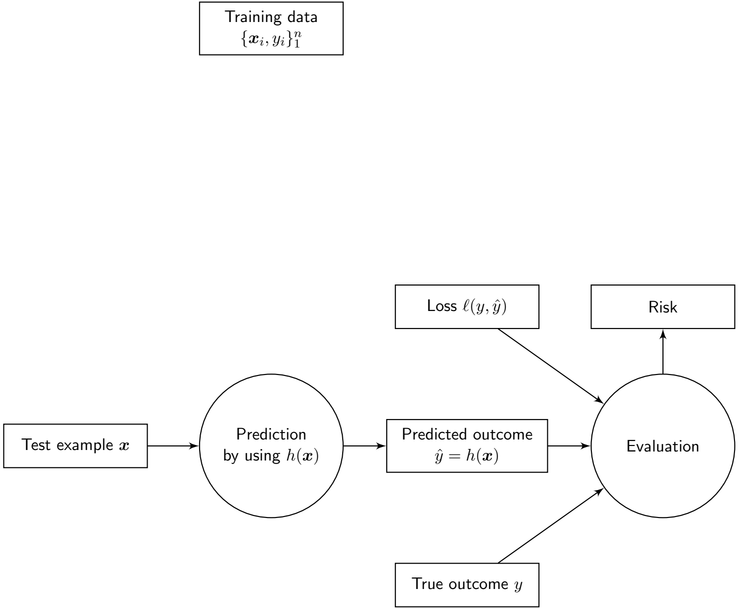
Supervised learning



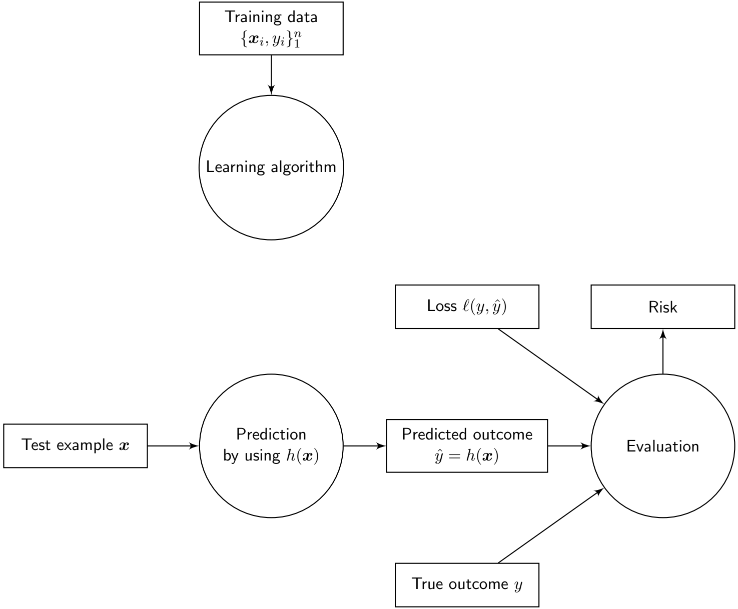
Supervised learning



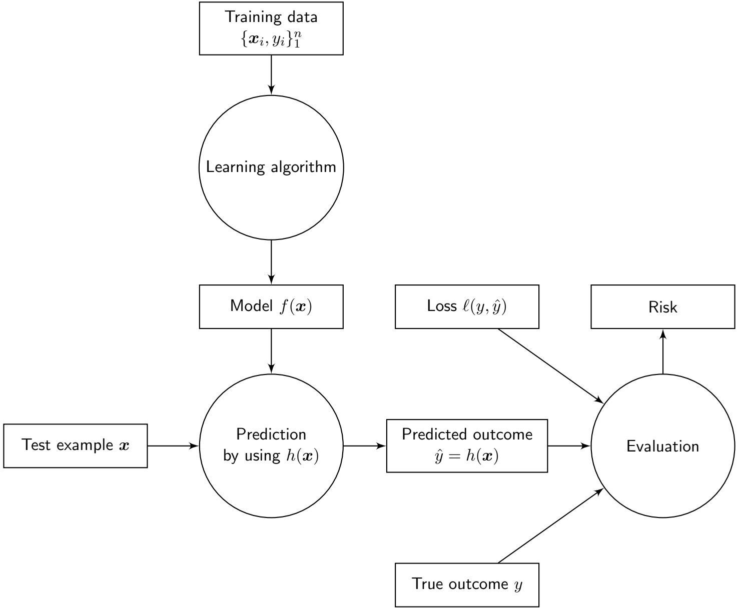
Supervised learning



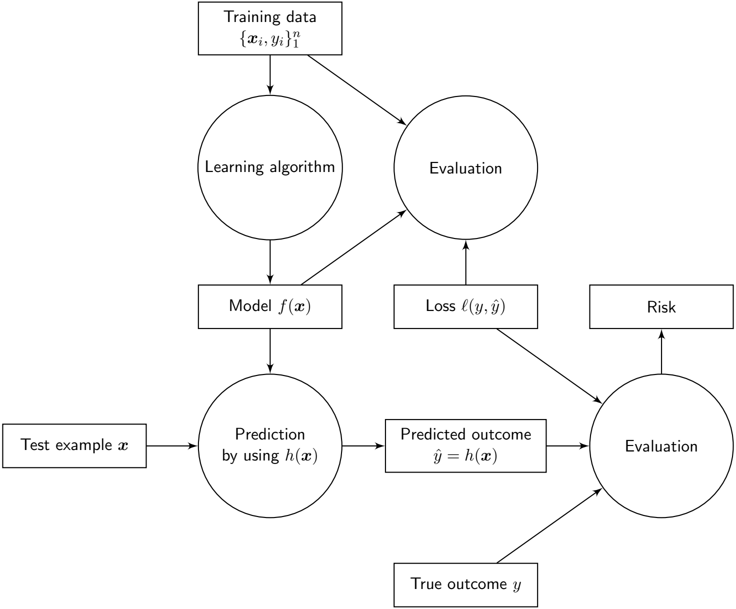
Supervised learning



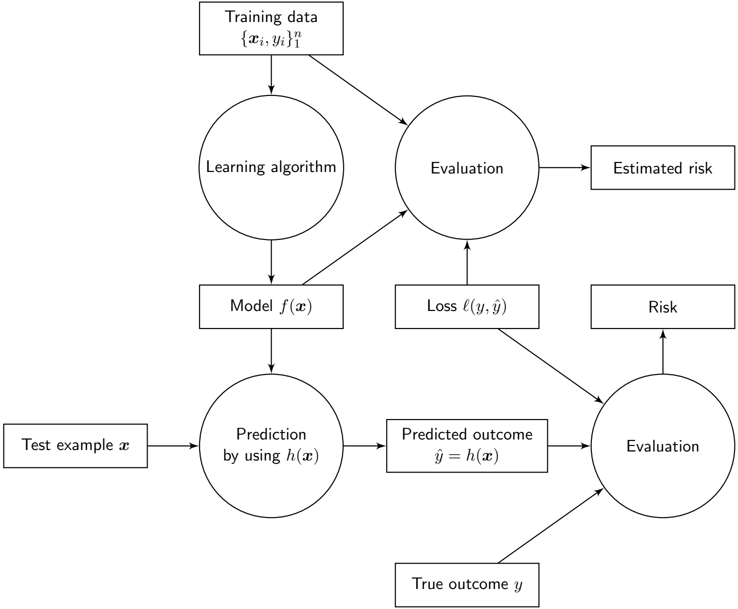
Supervised learning



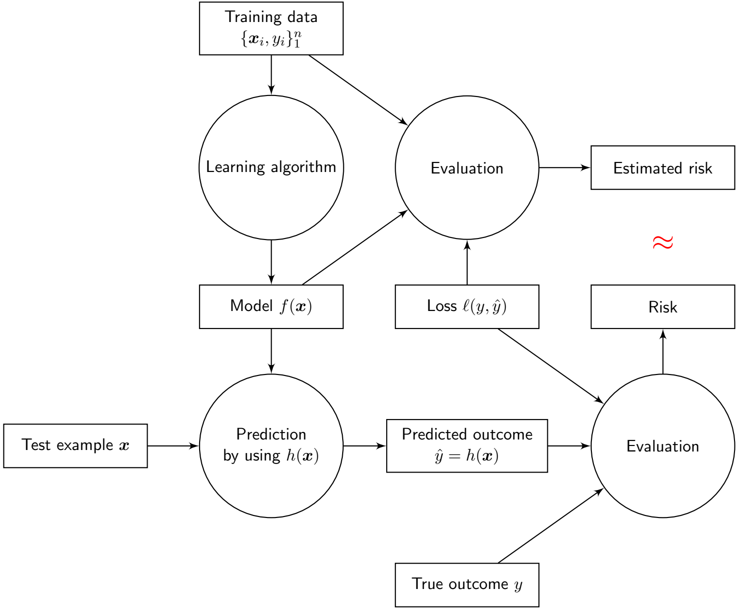
Supervised learning



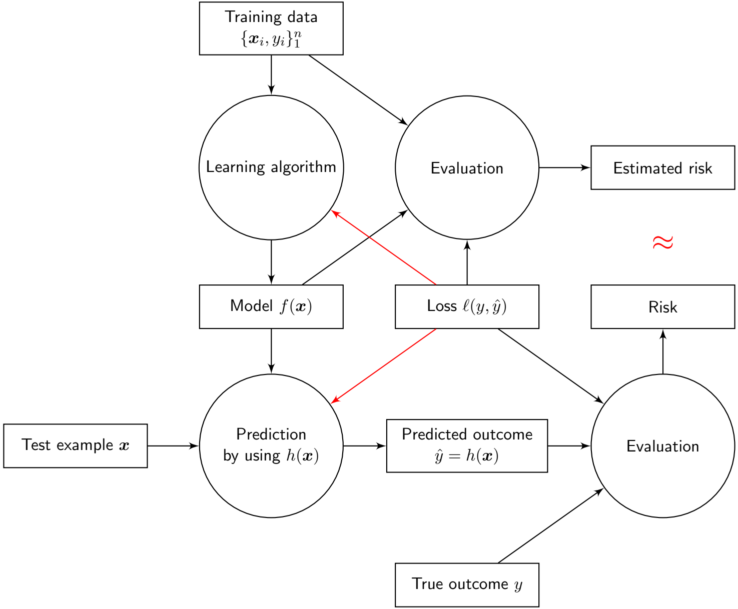
Supervised learning



Supervised learning



Supervised learning



Statistical learning framework

Statistical learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
  - usually a feature vector, $\mathcal{X} \subseteq \mathbb{R}^d$.

Statistical learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
  - usually a feature vector, $\mathcal{X} \subseteq \mathbb{R}^d$.
- **Outcome** $y \in \mathcal{Y}$ drawn from a distribution $P(y \mid \boldsymbol{x})$.
  - target of our prediction: class label, real value, label vector, etc.,
  - alternative view: **example** $(\boldsymbol{x}, y)$ drawn from $P(\boldsymbol{x}, y)$.

Statistical learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
  - usually a feature vector, $\mathcal{X} \subseteq \mathbb{R}^d$.
- **Outcome** $y \in \mathcal{Y}$ drawn from a distribution $P(y \,|\, \boldsymbol{x})$.
  - target of our prediction: class label, real value, label vector, etc.,
  - alternative view: **example** $(\boldsymbol{x}, y)$ drawn from $P(\boldsymbol{x}, y)$.
- **Prediction** $\hat{y} = h(\boldsymbol{x})$ by means of **prediction function** $h \colon \mathcal{X} \to \mathcal{Y}$.
  - h returns prediction $\hat{y} = h(\boldsymbol{x})$ for every input $\boldsymbol{x}$.

Statistical learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
  - usually a feature vector, $\mathcal{X} \subseteq \mathbb{R}^d$.
- **Outcome** $y \in \mathcal{Y}$ drawn from a distribution $P(y \,|\, \boldsymbol{x})$.
  - target of our prediction: class label, real value, label vector, etc.,
  - alternative view: **example** $(\boldsymbol{x}, y)$ drawn from $P(\boldsymbol{x}, y)$.
- **Prediction** $\hat{y} = h(\boldsymbol{x})$ by means of **prediction function** $h \colon \mathcal{X} \to \mathcal{Y}$.
  - h returns prediction $\hat{y} = h(\boldsymbol{x})$ for every input $\boldsymbol{x}$.
- **Loss** of our prediction: $\ell(y, \hat{y})$.
  - $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a problem-specific **loss function**.

Statistical learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
  - usually a feature vector, $\mathcal{X} \subseteq \mathbb{R}^d$.
- **Outcome** $y \in \mathcal{Y}$ drawn from a distribution $P(y \,|\, \boldsymbol{x})$.
  - target of our prediction: class label, real value, label vector, etc.,
  - alternative view: **example** $(\boldsymbol{x}, y)$ drawn from $P(\boldsymbol{x}, y)$.
- **Prediction** $\hat{y} = h(\boldsymbol{x})$ by means of **prediction function** $h \colon \mathcal{X} \to \mathcal{Y}$.
  - h returns prediction $\hat{y} = h(\boldsymbol{x})$ for every input $\boldsymbol{x}$.
- **Loss** of our prediction: $\ell(y, \hat{y})$.
  - $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a problem-specific **loss function**.
- **Goal**: find a prediction function with small loss.

Risk

- **Goal**: minimize the **expected** loss over all examples (**risk**):

$$L_\ell(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[\ell(y, h(\boldsymbol{x}))\right].$$

Risk

- **Goal**: minimize the **expected** loss over all examples (**risk**):

$$L_\ell(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[\ell(y, h(\boldsymbol{x}))\right].$$

- The **optimal** prediction function over all possible functions:

$$h^* = \arg\min_h L(h),$$

(so called **Bayes prediction function**).

Risk

- **Goal**: minimize the **expected** loss over all examples (**risk**):

$$L_\ell(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[\ell(y, h(\boldsymbol{x}))\right].$$

- The **optimal** prediction function over all possible functions:

$$h^* = \arg\min_h L(h),$$

  (so called **Bayes prediction function**).

- The smallest achievable risk (**Bayes risk**):

$$L_\ell^* = L_\ell(h^*).$$

Decomposition of risk

$$L_\ell(h) \quad = \quad \mathbb{E}_{(\boldsymbol{x},y)}\left[\ell(y, h(\boldsymbol{x}))\right]$$

Decomposition of risk

$$
\begin{aligned}
L_\ell(h) &= \mathbb{E}_{(\boldsymbol{x},y)}\left[\ell(y, h(\boldsymbol{x}))\right] \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)\mathrm{d}\boldsymbol{x}\mathrm{d}y
\end{aligned}
$$

Decomposition of risk

$$
\begin{aligned}
L_\ell(h) &= \mathbb{E}_{(\boldsymbol{x},y)}\left[\ell(y, h(\boldsymbol{x}))\right] \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(\boldsymbol{x})) P(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} \mathrm{d}y \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \ell(y, h(\boldsymbol{x})) P(y \mid \boldsymbol{x}) \mathrm{d}y \right) P(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}
\end{aligned}
$$

Decomposition of risk

$$
\begin{aligned}
L_\ell(h) &= \mathbb{E}_{(\boldsymbol{x},y)}\left[\ell(y, h(\boldsymbol{x}))\right] \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(\boldsymbol{x})) P(\boldsymbol{x}, y) \mathrm{d}\boldsymbol{x} \mathrm{d}y \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \ell(y, h(\boldsymbol{x})) P(y \,|\, \boldsymbol{x}) \mathrm{d}y \right) P(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{x}}\left[L_\ell(h \,|\, \boldsymbol{x})\right].
\end{aligned}
$$

Decomposition of risk

$$
\begin{aligned}
L_\ell(h) &= \mathbb{E}_{(\boldsymbol{x},y)}\left[\ell(y, h(\boldsymbol{x}))\right] \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)\mathrm{d}\boldsymbol{x}\mathrm{d}y \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \ell(y, h(\boldsymbol{x}))P(y \,|\, \boldsymbol{x})\mathrm{d}y \right) P(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{x}}\left[L_\ell(h \,|\, \boldsymbol{x})\right].
\end{aligned}
$$

- $L_\ell(h \,|\, \boldsymbol{x})$ is the **conditional risk** of $\hat{y} = h(\boldsymbol{x})$ at $\boldsymbol{x}$.

Decomposition of risk

$$\begin{aligned} L_\ell(h) &= \mathbb{E}_{(\boldsymbol{x},y)}\left[\ell(y, h(\boldsymbol{x}))\right] \\ &= \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(\boldsymbol{x}))P(\boldsymbol{x}, y)\mathrm{d}\boldsymbol{x}\mathrm{d}y \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \ell(y, h(\boldsymbol{x}))P(y \,|\, \boldsymbol{x})\mathrm{d}y \right) P(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \mathbb{E}_{\boldsymbol{x}}\left[L_\ell(h \,|\, \boldsymbol{x})\right]. \end{aligned}$$

- $L_\ell(h \,|\, \boldsymbol{x})$ is the **conditional risk** of $\hat{y} = h(\boldsymbol{x})$ at $\boldsymbol{x}$.
- Bayes prediction **minimizes the conditional risk** for every $\boldsymbol{x}$:

$$h^*(\boldsymbol{x}) = \arg\min_h L_\ell(h \,|\, \boldsymbol{x}).$$

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).

Making optimal decisions

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).
- Decision = bet (four choices).

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).
- Decision = bet (four choices).
- If you win you get 100$, if you loose you must give 50$.

Making optimal decisions

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).
- Decision = bet (four choices).
- If you win you get 100$, if you loose you must give 50$.
- What is the loss and optimal decision?

Making optimal decisions

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).
- Decision = bet (four choices).
- If you win you get 100$, if you loose you must give 50$.
- What is the loss and optimal decision?
- Suppose we know the card is black. What is the optimal decision now?

Making optimal decisions

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).
- Decision = bet (four choices).
- If you win you get 100$, if you loose you must give 50$.
- What is the loss and optimal decision?
- Suppose we know the card is black. What is the optimal decision now?
- What are the input variables?

Making optimal decisions

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).
- Bet the color:

Making optimal decisions

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).
- Bet the color:
  - if the true color is red and you are correct you win 50, otherwise you loose 100,

Making optimal decisions

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).
- Bet the color:
  - if the true color is red and you are correct you win 50, otherwise you loose 100,
  - if the true color is black and you are correct you win 200, otherwise you loose 100.

Example

- Pack of cards: 7 diamonds (red), 5 hearts (red), 5 spades (black), 2 clubs (black).
- Bet the color:
  - if the true color is red and you are correct you win 50, otherwise you loose 100,
  - if the true color is black and you are correct you win 200, otherwise you loose 100.
- What is the loss and optimal decision now?

Regression

- Prediction of a **real-valued** outcome $y \in \mathbb{R}$.
- Find a prediction function $h(\boldsymbol{x})$ that accurately predicts value of y.
- The most common loss function used is **squared error loss**:

$$\ell_{se}(y, \hat{y}) = (y - \hat{y})^2,$$

where $\hat{y} = h(\boldsymbol{x})$.

Regression

- The conditional risk for squared error loss is :

Regression

- The conditional risk for squared error loss is :

$$L_{se}(h \,|\, \boldsymbol{x}) = \mathbb{E}_{y|\boldsymbol{x}} \left[(y - \hat{y})^2 \right]$$

Regression

- The conditional risk for squared error loss is :

$\mu(\boldsymbol{x}) = \mathbb{E}_{y \mid \boldsymbol{x}}[y]$

$$
\begin{aligned}
L_{se}(h \mid \boldsymbol{x}) &= \mathbb{E}_{y|\boldsymbol{x}}\left[(y - \hat{y})^2\right] \\
&= \mathbb{E}_{y|\boldsymbol{x}}\left[(y - \mu(\boldsymbol{x}) + \mu(\boldsymbol{x}) - \hat{y})^2\right]
\end{aligned}
$$

- The conditional risk for squared error loss is :

$\mu(\boldsymbol{x}) = \mathbb{E}_{y \mid \boldsymbol{x}}[y]$

$$L_{se}(h \mid \boldsymbol{x}) = \mathbb{E}_{y|\boldsymbol{x}}\left[(y-\hat{y})^2\right]$$
$$= \mathbb{E}_{y|\boldsymbol{x}}\left[(y-\mu(\boldsymbol{x})+\mu(\boldsymbol{x})-\hat{y})^2\right]$$
$$= \mathbb{E}_{y|\boldsymbol{x}}\Big[(y-\mu(\boldsymbol{x}))^2 + 2\underbrace{(y-\mu(\boldsymbol{x}))}_{=0 \text{ under expectation}}(\mu(\boldsymbol{x})-\hat{y}) + (\mu(\boldsymbol{x})-\hat{y})^2\Big]$$

- The conditional risk for squared error loss is :

$\mu(\boldsymbol{x}) = \mathbb{E}_{y \mid \boldsymbol{x}}[y]$

$$
\begin{aligned}
L_{se}(h \mid \boldsymbol{x}) &= \mathbb{E}_{y|\boldsymbol{x}}\left[(y-\hat{y})^2\right] \\
&= \mathbb{E}_{y|\boldsymbol{x}}\left[(y-\mu(\boldsymbol{x})+\mu(\boldsymbol{x})-\hat{y})^2\right] \\
&= \mathbb{E}_{y|\boldsymbol{x}}\Big[(y-\mu(\boldsymbol{x}))^2 + 2\underbrace{(y-\mu(\boldsymbol{x}))}_{=0 \text{ under expectation}}(\mu(\boldsymbol{x})-\hat{y}) + (\mu(\boldsymbol{x})-\hat{y})^2\Big] \\
&= \underbrace{\mathbb{E}_{y|\boldsymbol{x}}\left[(y-\mu(\boldsymbol{x}))^2\right]}_{\text{independent of } \hat{y}} + (\mu(\boldsymbol{x})-\hat{y})^2.
\end{aligned}
$$

- The conditional risk for squared error loss is :

$\mu(\boldsymbol{x}) = \mathbb{E}_{y \mid \boldsymbol{x}}[y]$

$$
\begin{aligned}
L_{se}(h \mid \boldsymbol{x}) &= \mathbb{E}_{y|\boldsymbol{x}}\left[(y - \hat{y})^2\right] \\
&= \mathbb{E}_{y|\boldsymbol{x}}\left[(y - \mu(\boldsymbol{x}) + \mu(\boldsymbol{x}) - \hat{y})^2\right] \\
&= \mathbb{E}_{y|\boldsymbol{x}}\Big[(y - \mu(\boldsymbol{x}))^2 + 2\underbrace{(y - \mu(\boldsymbol{x}))}_{=0 \text{ under expectation}}(\mu(\boldsymbol{x}) - \hat{y}) + (\mu(\boldsymbol{x}) - \hat{y})^2\Big] \\
&= \underbrace{\mathbb{E}_{y|\boldsymbol{x}}\left[(y - \mu(\boldsymbol{x}))^2\right]}_{\text{independent of } \hat{y}} + (\mu(\boldsymbol{x}) - \hat{y})^2.
\end{aligned}
$$

- Hence, $h^*(\boldsymbol{x}) = \mu(\boldsymbol{x})$, the **conditional expectation** of y at $\boldsymbol{x}$, and:

$$L_{se}(h^* \mid \boldsymbol{x}) = \mathbb{E}_{y|\boldsymbol{x}}\left[(y - \mu(\boldsymbol{x}))^2\right] = \mathrm{Var}(y|\boldsymbol{x}).$$

Regression

- Another loss commonly used in regression is the **absolute error**:

$$\ell_{ae}(y, \hat{y}) = |y - \hat{y}|.$$

- The Bayes classifier for the absolute-error loss is:

$$h^*(\boldsymbol{x}) = \arg\min_h L_{ae}(h \,|\, \boldsymbol{x}) =$$

Regression

- Another loss commonly used in regression is the **absolute error**:

$$\ell_{ae}(y, \hat{y}) = |y - \hat{y}|.$$

- The Bayes classifier for the absolute-error loss is:

$$h^*(\boldsymbol{x}) = \arg\min_h L_{ae}(h \,|\, \boldsymbol{x}) = \mathrm{median}(y|\boldsymbol{x})\,,$$

i.e., **median** of the conditional distribution of y given $\boldsymbol{x}$.

Regression

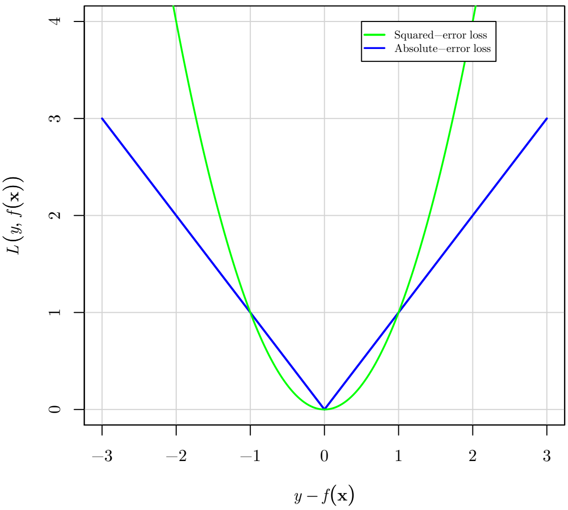


Figure: Loss functions for regression task

Binary Classification

- Prediction of a **binary** outcome $y \in \{-1, 1\}$ (alternatively $y \in \{0, 1\}$).
- Find a prediction function $h(\boldsymbol{x})$ that accurately predicts value of y.
- The most common loss function used is **0/1 loss**:

$$\ell_{0/1}(y, \hat{y}) = [\![y \neq \hat{y}]\!] = \begin{cases} 0, & \text{if } y = \hat{y}\,, \\ 1, & \text{otherwise}\,. \end{cases}$$

Binary Classification

- Define $\eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$.

Binary Classification

- Define $\eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$.
- The conditional 0/1 risk at $\boldsymbol{x}$ is:

Binary Classification

- Define $\eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$.
- The conditional 0/1 risk at $\boldsymbol{x}$ is:

$$L_{0/1}(h|\boldsymbol{x}) = \eta(\boldsymbol{x})[\![h(\boldsymbol{x}) \neq 1]\!] + (1 - \eta(\boldsymbol{x}))[\![h(\boldsymbol{x}) \neq -1]\!].$$

- Define $\eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$.
- The conditional 0/1 risk at $\boldsymbol{x}$ is:

$$L_{0/1}(h|\boldsymbol{x}) = \eta(\boldsymbol{x})[\![h(\boldsymbol{x}) \neq 1]\!] + (1 - \eta(\boldsymbol{x}))[\![h(\boldsymbol{x}) \neq -1]\!].$$

- The **Bayes classifier**:

Binary Classification

- Define $\eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$.
- The conditional 0/1 risk at $\boldsymbol{x}$ is:

$$L_{0/1}(h|\boldsymbol{x}) = \eta(\boldsymbol{x})[\![h(\boldsymbol{x}) \neq 1]\!] + (1 - \eta(\boldsymbol{x}))[\![h(\boldsymbol{x}) \neq -1]\!].$$

- The **Bayes classifier**:

$$h^*(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \eta(\boldsymbol{x}) > 1 - \eta(\boldsymbol{x}) \\ -1 & \text{if } \eta(\boldsymbol{x}) < 1 - \eta(\boldsymbol{x}) \end{cases} = \operatorname{sgn}\left(\eta(\boldsymbol{x}) - 1/2\right),$$

and the **Bayes conditional risk**:

- Define $\eta(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$.
- The conditional 0/1 risk at $\boldsymbol{x}$ is:

$$L_{0/1}(h|\boldsymbol{x}) = \eta(\boldsymbol{x})[\![h(\boldsymbol{x}) \neq 1]\!] + (1 - \eta(\boldsymbol{x}))[\![h(\boldsymbol{x}) \neq -1]\!].$$

- The **Bayes classifier**:

$$h^*(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \eta(\boldsymbol{x}) > 1 - \eta(\boldsymbol{x}) \\ -1 & \text{if } \eta(\boldsymbol{x}) < 1 - \eta(\boldsymbol{x}) \end{cases} = \operatorname{sgn}\left(\eta(\boldsymbol{x}) - 1/2\right),$$

and the **Bayes conditional risk**:

$$L_\ell(h^* \,|\, \boldsymbol{x}) = \min\{\eta(\boldsymbol{x}), 1 - \eta(\boldsymbol{x})\}.$$

Multi-class classification

- **Domain** of outcome variable y is a set of labels $\mathcal{Y} = \{1, \ldots, K\}$.
- **Goal**: find a prediction function $h(\boldsymbol{x})$ that for any object $\boldsymbol{x}$ predicts accurately the actual value of y.
- **Loss function**: the most common is 0/1 loss:

$$\ell_{0/1}(y, \hat{y}) = \begin{cases} 0, & \text{if } y = \hat{y}\,, \\ 1, & \text{otherwise}\,. \end{cases}$$

Multi-class classification

- The conditional risk of the 0/1 loss is:

$$L_{0/1}(h \mid \boldsymbol{x}) \quad = \quad \mathbb{E}_{y|\boldsymbol{x}} \ell_{0/1}(y, h(\boldsymbol{x}))$$

Multi-class classification

- The conditional risk of the 0/1 loss is:

$$\begin{aligned} L_{0/1}(h \,|\, \boldsymbol{x}) &= \mathbb{E}_{y|\boldsymbol{x}} \ell_{0/1}(y, h(\boldsymbol{x})) \\ &= \sum_{k \in \mathcal{Y}} P(y = k|\boldsymbol{x}) \ell_{0/1}(k, h(\boldsymbol{x})) \end{aligned}$$

Multi-class classification

- The conditional risk of the 0/1 loss is:

$$\begin{aligned} L_{0/1}(h \,|\, \boldsymbol{x}) &= \mathbb{E}_{y|\boldsymbol{x}} \ell_{0/1}(y, h(\boldsymbol{x})) \\ &= \sum_{k \in \mathcal{Y}} P(y = k|\boldsymbol{x}) \ell_{0/1}(k, h(\boldsymbol{x})) \end{aligned}$$

- Therefore, the Bayes classifier is given by:

$$h^*(\boldsymbol{x}) = \arg\min_{h} L_{0/1}(h \,|\, \boldsymbol{x})$$

Multi-class classification

- The conditional risk of the 0/1 loss is:

$$\begin{aligned} L_{0/1}(h \,|\, \boldsymbol{x}) &= \mathbb{E}_{y|\boldsymbol{x}} \ell_{0/1}(y, h(\boldsymbol{x})) \\ &= \sum_{k \in \mathcal{Y}} P(y = k|\boldsymbol{x}) \ell_{0/1}(k, h(\boldsymbol{x})) \end{aligned}$$

- Therefore, the Bayes classifier is given by:

$$\begin{aligned} h^*(\boldsymbol{x}) &= \arg\min_h L_{0/1}(h \,|\, \boldsymbol{x}) \\ &= \arg\max_k P(y = k|\boldsymbol{x}) \,, \end{aligned}$$

the class with **the largest conditional probability** $P(y|\boldsymbol{x})$.

Multi-class classification

- The conditional risk of the 0/1 loss is:

$$
\begin{aligned}
L_{0/1}(h \,|\, \boldsymbol{x}) &= \mathbb{E}_{y|\boldsymbol{x}} \ell_{0/1}(y, h(\boldsymbol{x})) \\
&= \sum_{k \in \mathcal{Y}} P(y = k|\boldsymbol{x}) \ell_{0/1}(k, h(\boldsymbol{x}))
\end{aligned}
$$

- Therefore, the Bayes classifier is given by:

$$
\begin{aligned}
h^*(\boldsymbol{x}) &= \arg\min_h L_{0/1}(h \,|\, \boldsymbol{x}) \\
&= \arg\max_k P(y = k|\boldsymbol{x}) \,,
\end{aligned}
$$

  the class with **the largest conditional probability** $P(y|\boldsymbol{x})$.
- The Bayes conditional risk:

Multi-class classification

- The conditional risk of the 0/1 loss is:

$$\begin{aligned} L_{0/1}(h \,|\, \boldsymbol{x}) &= \mathbb{E}_{y|\boldsymbol{x}} \ell_{0/1}(y, h(\boldsymbol{x})) \\ &= \sum_{k \in \mathcal{Y}} P(y = k|\boldsymbol{x}) \ell_{0/1}(k, h(\boldsymbol{x})) \end{aligned}$$

- Therefore, the Bayes classifier is given by:

$$\begin{aligned} h^*(\boldsymbol{x}) &= \arg\min_h L_{0/1}(h \,|\, \boldsymbol{x}) \\ &= \arg\max_k P(y = k|\boldsymbol{x}), \end{aligned}$$

  the class with **the largest conditional probability** $P(y|\boldsymbol{x})$.
- The Bayes conditional risk:

$$L_\ell(h^* \,|\, \boldsymbol{x}) = \min\{1 - P(y = k|\boldsymbol{x}) : k \in \mathcal{Y}\}.$$

Deterministic learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
- **Outcome** $y \in \mathcal{Y}$.
- **Unknown target function** $h^*\colon \mathcal{X} \to \mathcal{Y}$, such that $y = h^*(\boldsymbol{x})$.
- **Goal**: discover h^* by observing examples of $(\boldsymbol{x}, y)$.

Deterministic learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
- **Outcome** $y \in \mathcal{Y}$.
- **Unknown target function** $h^*\colon \mathcal{X} \to \mathcal{Y}$, such that $y = h^*(\boldsymbol{x})$.
- **Goal**: discover h^* by observing examples of $(\boldsymbol{x}, y)$.

- This is a **special case** of the statistical framework:
  - What is $P(y|\boldsymbol{x})$?
  - Bayes prediction function?
  - Risk of h^*? (assuming $\ell(y, \hat{y}) = 0$ whenever $y = \hat{y}$)

Deterministic learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
- **Outcome** $y \in \mathcal{Y}$.
- **Unknown target function** $h^*\colon \mathcal{X} \to \mathcal{Y}$, such that $y = h^*(\boldsymbol{x})$.
- **Goal**: discover h^* by observing examples of $(\boldsymbol{x}, y)$.

- This is a **special case** of the statistical framework:
  - What is $P(y|\boldsymbol{x})$?
    - $P(y|\boldsymbol{x})$ is a **degenerate** distribution for every $\boldsymbol{x}$.
  - Bayes prediction function?
  - Risk of h^*? (assuming $\ell(y, \hat{y}) = 0$ whenever $y = \hat{y}$)

Deterministic learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
- **Outcome** $y \in \mathcal{Y}$.
- **Unknown target function** $h^*: \mathcal{X} \to \mathcal{Y}$, such that $y = h^*(\boldsymbol{x})$.
- **Goal**: discover h^* by observing examples of $(\boldsymbol{x}, y)$.

- This is a **special case** of the statistical framework:
  - What is $P(y|\boldsymbol{x})$?
    - $P(y|\boldsymbol{x})$ is a **degenerate** distribution for every $\boldsymbol{x}$.
  - Bayes prediction function?
    - h^*
  - Risk of h^*? (assuming $\ell(y, \hat{y}) = 0$ whenever $y = \hat{y}$)

Deterministic learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
- **Outcome** $y \in \mathcal{Y}$.
- **Unknown target function** $h^*\colon \mathcal{X} \to \mathcal{Y}$, such that $y = h^*(\boldsymbol{x})$.
- **Goal**: discover h^* by observing examples of $(\boldsymbol{x}, y)$.

- This is a **special case** of the statistical framework:
  - What is $P(y|\boldsymbol{x})$?
    - $P(y|\boldsymbol{x})$ is a **degenerate** distribution for every $\boldsymbol{x}$.
  - Bayes prediction function?
    - h^*
  - Risk of h^*? (assuming $\ell(y, \hat{y}) = 0$ whenever $y = \hat{y}$)
    - h^* has **zero risk**.

Deterministic learning framework

- **Input** $\boldsymbol{x} \in \mathcal{X}$ drawn from a distribution $P(\boldsymbol{x})$.
- **Outcome** $y \in \mathcal{Y}$.
- **Unknown target function** $h^*\colon \mathcal{X} \to \mathcal{Y}$, such that $y = h^*(\boldsymbol{x})$.
- **Goal**: discover h^* by observing examples of $(\boldsymbol{x}, y)$.

- This is a **special case** of the statistical framework:
  - What is $P(y|\boldsymbol{x})$?
    - $P(y|\boldsymbol{x})$ is a **degenerate** distribution for every $\boldsymbol{x}$.
  - Bayes prediction function?
    - h^*
  - Risk of h^*? (assuming $\ell(y, \hat{y}) = 0$ whenever $y = \hat{y}$)
    - h^* has **zero risk**.
  - Unrealistic scenario in real life.

Outline

Learning

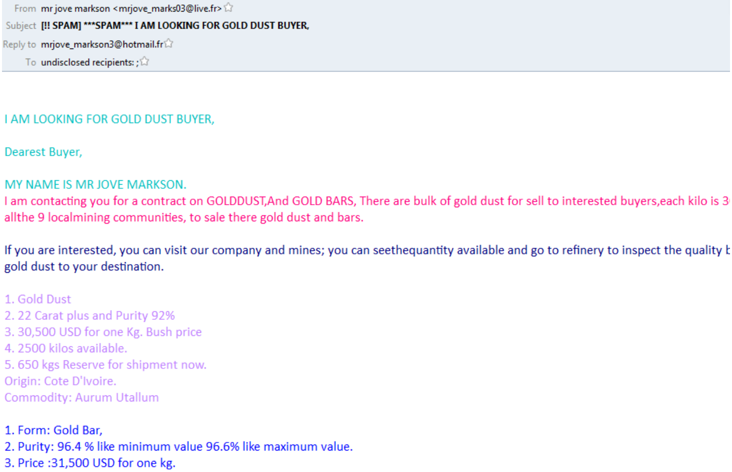
- Distribution $P(\boldsymbol{x}, y)$ is unknown **unknown**.
- Therefore, Bayes classifier h^* is also **unknown**.
- Instead, we have access to n independent and identically distributed (i.i.d) **training examples** (**sample**):

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}.$$

- **Learning**: use training data to find a good **approximation** of h^*.

Spam filtering

- Problem: Predict whether a given email is spam or not.
- An object to be classified: an email.
- There are two possible responses (classes): spam, not spam.



From mr jove markson <mrjove_marks03@live.fr>
Subject [!! SPAM] ***SPAM*** I AM LOOKING FOR GOLD DUST BUYER,
Reply to mrjove_markson3@hotmail.fr
To undisclosed recipients: ;

I AM LOOKING FOR GOLD DUST BUYER,

Dearest Buyer,

MY NAME IS MR JOVE MARKSON.
I am contacting you for a contract on GOLDDUST,And GOLD BARS, There are bulk of gold dust for sell to interested buyers,each kilo is 3
allthe 9 localmining communities, to sale there gold dust and bars.

If you are interested, you can visit our company and mines; you can seethequantity available and go to refinery to inspect the quality b
gold dust to your destination.

1. Gold Dust
2. 22 Carat plus and Purity 92%
3. 30,500 USD for one Kg. Bush price
4. 2500 kilos available.
5. 650 kgs Reserve for shipment now.
Origin: Cote D'Ivoire.
Commodity: Aurum Utallum

1. Form: Gold Bar,
2. Purity: 96.4 % like minimum value 96.6% like maximum value.
3. Price :31,500 USD for one kg.

Spam filtering

Example

- Representation of an email through (meaningful) features:

Spam filtering

Example

- Representation of an email through (meaningful) features:
  - length of subject
  - length of email body,
  - use of colors,
  - domain,
  - words in subject,
  - words in body.

length of subject	length of body	use of colors	domain	gold	price	USD	...	machine	learning	spam?
7	240	1	live.fr	1	1	1	...	0	0	1
2	150	0	poznan.pl	0	0	0	...	1	1	0
2	250	0	tibco.com	0	1	1	...	1	1	0
4	120	1	r-project.org	0	1	0	...	0	0	?

Training/Test Data in Computer Format

Example (ARFF format for training/test data)

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {true, false}
@attribute play {yes, no}

@data
sunny,85,85,false,no
sunny,80,90,true,no
overcast,83,86,false,yes
rainy,70,96,false,yes
rainy,68,80,false,yes
rainy,65,70,true,no
overcast,64,65,true,yes
sunny,72,95,false,no
sunny,69,70,false,yes
rainy,75,80,false,yes
sunny,75,70,true,yes
overcast,72,90,true,yes
overcast,81,75,true,yes
rainy,71,91,true,no
```

Learning

- Four types of datasets:
  - **training** data: past emails,
  - **validation** data: a portion of past email used for tuning learning algorithms
  - **test** data: a portion of past emails used for estimating the risk,
  - **new incoming** data to be classified: new incoming emails.

Different learning paradigms

Different learning paradigms

- **Generative learning**

Different learning paradigms

- **Generative learning**
  - Follow a data generating process
  - Learn a model of the joint distribution $P(\boldsymbol{x}, y)$ and then use the Bayes theorem to obtain $P(y \mid \boldsymbol{x})$.
  - Make the final prediction by computing the optimal decision based on $P(y \mid \boldsymbol{x})$ with respect to a given $\ell(y, \hat{y})$.

Different learning paradigms

- **Generative learning**
  - Follow a data generating process
  - Learn a model of the joint distribution $P(\boldsymbol{x}, y)$ and then use the Bayes theorem to obtain $P(y \mid \boldsymbol{x})$.
  - Make the final prediction by computing the optimal decision based on $P(y \mid \boldsymbol{x})$ with respect to a given $\ell(y, \hat{y})$.
- **Discriminative learning**

Different learning paradigms

- **Generative learning**
  - Follow a data generating process
  - Learn a model of the joint distribution $P(\boldsymbol{x}, y)$ and then use the Bayes theorem to obtain $P(y \mid \boldsymbol{x})$.
  - Make the final prediction by computing the optimal decision based on $P(y \mid \boldsymbol{x})$ with respect to a given $\ell(y, \hat{y})$.
- **Discriminative learning**
  - Approximate $h^*(\boldsymbol{x})$ which is a direct map from $\boldsymbol{x}$ to y or
  - Model the conditional probability $P(y \mid \boldsymbol{x})$ directly, and
  - Make the final prediction by computing the optimal decision based on $P(y \mid \boldsymbol{x})$ with respect to a given $\ell(y, \hat{y})$.

Different learning paradigms

- **Generative learning**
  - Follow a data generating process
  - Learn a model of the joint distribution $P(\boldsymbol{x}, y)$ and then use the Bayes theorem to obtain $P(y \mid \boldsymbol{x})$.
  - Make the final prediction by computing the optimal decision based on $P(y \mid \boldsymbol{x})$ with respect to a given $\ell(y, \hat{y})$.
- **Discriminative learning**
  - Approximate $h^*(\boldsymbol{x})$ which is a direct map from $\boldsymbol{x}$ to y or
  - Model the conditional probability $P(y \mid \boldsymbol{x})$ directly, and
  - Make the final prediction by computing the optimal decision based on $P(y \mid \boldsymbol{x})$ with respect to a given $\ell(y, \hat{y})$.
- **Two phases** of the learning models: learning and prediction (inference).

Different learning paradigms

- Various principles on how to learn:
  - Empirical risk minimization,
  - Maximum likelihood principle,
  - Bayes approach,
  - Minimum description length,
  - ...

Empirical Risk Minimization (ERM)

- Choose a prediction function $\widehat{h}$ which minimizes the loss on the training data within some **restricted** class of functions $\mathcal{H}$.

$$\widehat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(\boldsymbol{x}_i)).$$

- The average loss on the training data is called **empirical risk** $\widehat{L}_\ell(h)$.

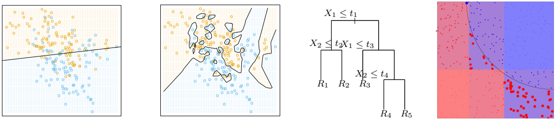
1

Empirical Risk Minimization (ERM)

- Choose a prediction function $\widehat{h}$ which minimizes the loss on the training data within some **restricted** class of functions $\mathcal{H}$.

$$\widehat{h} = \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, h(\boldsymbol{x}_i)).$$

- The average loss on the training data is called **empirical risk** $\widehat{L}_\ell(h)$.
- $\mathcal{H}$ can be: linear functions, polynomials, trees of a given depth, rules, linear combinations of trees, etc.[1]



[1] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Second Edition*. Springer, 2009

Outline

Summary

- What is machine learning?
- Supervised learning: statistical decision/learning theory, loss functions, risk.
- Learning paradigms and principles.

Bibliography

- T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Second Edition*.
  Springer, 2009
  http://www-stat.stanford.edu/~tibs/ElemStatLearn/

- Christopher M. Bishop. *Pattern Recognition and Machine Learning*.
  Springer-Verlag, 2006

- David Barber. *Bayesian Reasoning and Machine Learning*.
  Cambridge University Press, 2012
  http://www.cs.ucl.ac.uk/staff/d.barber/brml/

- Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*.
  AMLBook, 2012