Lecture 2: Bayesian Decision Theory

Isabel Valera

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

22.04.2023

Outline

1 Bibliography

2 Introduction

3 Bayes rule

4 Bayes classifier

5 Cost-sensitive class.

6 Margin-based class.

7 Multi-class class.

8 Regression

Main references

- Duda, Hart & Stork (DHS) - Chapter 2
- Bishop - Chapter 1.5

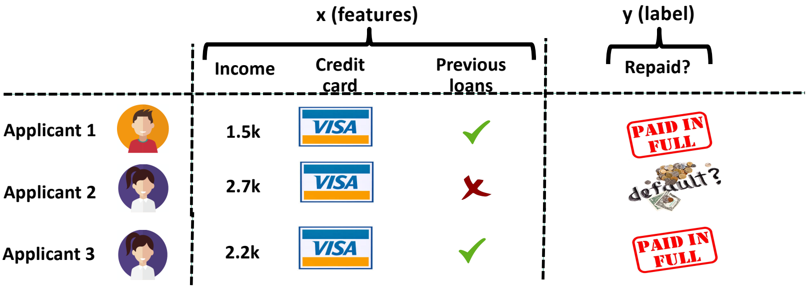
Outline

1. Bibliography
2. Introduction
3. Bayes rule
4. Bayes classifier
5. Cost-sensitive class.
6. Margin-based class.
7. Multi-class class.
8. Regression

Bayesian decision theory

Bayesian decision theory addresses the problem of making *optimal decisions under uncertainty*.

- **A decision rule** prescribes what decision to make based on observed input (e.g., grant the credit).
- **Uncertainty**: Usually Y is not a deterministic function of X but instead we assume a probability distribution $P(Y|X)$ that determines the probability of observing class $Y = y$ for the given features $X = x$.



	x (features)			y (label)
	Income	Credit card	Previous loans	Repaid?
Applicant 1	1.5k	VISA	✓	PAID IN FULL
Applicant 2	2.7k	VISA	✗	default?
Applicant 3	2.2k	VISA	✓	PAID IN FULL

Notation I

We denote by:

- X and Y the random variables corresponding to the features and the label, respectively.
- $\mathcal{X}$ and $\mathcal{Y}$ the sample spaces of respectively X and Y, e.g., $\mathcal{X} = \Re$ and $\mathcal{Y} = \{-1, 1\}$.
- x and y a concrete value taken by, respectively, the random variables X and Y.

Notation II

Let's for now assume $\mathcal{Y} = \{-1, 1\}$ and $p(X = x, Y = y)$ denotes the **joint density** of the probability measure P on $\mathcal{X} \times \mathcal{Y}$, which satisfies that:

$$P(Y = y|X = x) = \frac{p(X = x|Y = y) \times P(Y = y)}{p(X = x)},$$

where

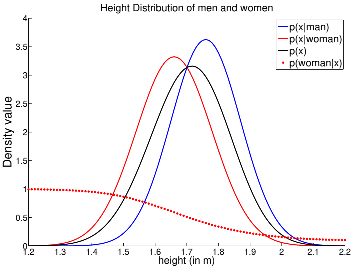
- $P(Y = y|X = x)$ denotes the **posterior probability** and corresponds to the probability that we observe y after observing x.
- $p(X = x|Y = y)$ denotes the **class-conditional density (or likelihood)** and models the occurrence of the features x of class y.
- $P(Y = y)$ denotes the **prior probability** of a class y and reflects our knowledge of how likely we expect a certain class before we can actually observe any data.
- $p(X = x)$ denotes the **marginal density (or evidence)** of the features x and models the cumulated occurrence of features over all classes $y \in \mathcal{Y}$.

Note: From now on we will denote $P(Y = y, X = x)$ by $P(y, x)$ to avoid a clutered notation.

Example I

Goal: Predict sex of a person (i.e., $\mathcal{Y} = \{\text{male}, \text{female}\}$) using height as feature (i.e., $\mathcal{X} = \mathbb{R}$). How do we find the optimal **classification rule**? 3 options:

1. Based on prior knowledge, i.e., classify x as female if $P(\text{female}) \geq P(\text{male})$.
2. Based on class conditional density, i.e., classify x as female if $p(x|\text{female}) \geq p(x|\text{male})$.
3. Based on posterior probability, i.e., classify x as female if $P(\text{female}|x) \geq P(\text{male}|x)$.



Example I

Goal: Predict sex of a person (i.e., $\mathcal{Y} = \{\text{male}, \text{female}\}$) using height as feature (i.e., $\mathcal{X} = \mathbb{R}$). How do we find the optimal **classification rule**? 3 options:

1. Based on prior knowledge, i.e., classify x as female if $P(\text{female}) \geq P(\text{male})$.
2. Based on class conditional density, i.e., classify x as female if $p(x|\text{female}) \geq p(x|\text{male})$.
3. Based on posterior probability, i.e., classify x as female if $P(\text{female}|x) \geq P(\text{male}|x)$.

→ Always decides same class for all x. $P(error|x) = P(error) = \min[P(male), P(female)]$.

→ For an observed feature vector x, $P(error|x) = \min[p(x|male), p(x|female)]$.

→ For an observed feature vector x, $P(error|x) = \min[P(male|x), P(female|x)]$.

Example II

Goal: Predict type of fish (i.e., $\mathcal{Y} = \{\omega_1, \omega_2\}$) using a set of features (i.e., $\mathcal{X} = \mathbb{R}^d$) such as length, width, lightness, etc.

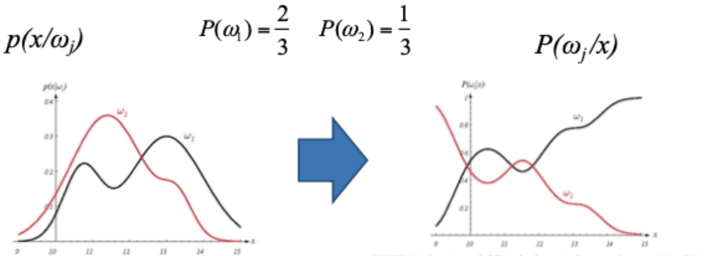


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons,

FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Figure: Images from DHS

Optimal decision

The optimal decision rule is given by:

$$y^* = \arg\max_{y_i \in \mathcal{Y}} P(Y = y_i|x),$$

is optimal, i.e., it minimizes $P(error|x)$ for all x and thus $P(error)$, which are given respectively by:

$$P(error|x) = \min[P(y_1|x), P(y_2|x)] \text{ (in binary cases)},$$

and

$$P(error) = \int P(error|x)p(x)dx.$$

It minimizes $P(error|x)$ for all x and thus also $P(error)$.

Outline

Learning to decide

- In machine learning, we often map decision problems to prediction problems, e.g., classification or regression problems.
- In both cases, we aim to learn a function, a.k.a, learning rule $\hat{y} : \mathcal{X} \to \mathcal{Y}$.
- We want to select the **Bayes optimal learning rule** that minimize the error, which we measure using a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ that computes (a proxy of) the error between the true and the predictive target values, $L(y, \hat{y}(x))$.
- Next, we show how to define the loss fucntion as well as how to derive the Bayes optimal learning rule for any classification or regression problem.

Loss function and risk

We first need to define a **quantitative measure of error:**

Definition (Loss function)

A **loss function** L is a mapping $L : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$.

Definition (Risk)

The **risk** or **expected loss** of a learning rule $\hat{y} : \mathcal{X} \to \mathcal{Y}$ is defined as

$$R_L(\hat{y}) = \mathbb{E}\big[L(\hat{y}(X), Y)\big] = \mathbb{E}\big[\mathbb{E}[L(\hat{y}(X), Y)|X]\big].$$

Note: $\mathbb{E}\big[\mathbb{E}[L(\hat{y}(X), Y)|X]\big] \;= \int_{\mathcal{X}} \Big[\int_{\mathcal{Y}} L(\hat{y}(x), y)\, p(y|x) dy \Big]\, p(x)\, dx.$

Bayes optimal risk

Definition

The **Bayes optimal risk** is given by

$$R_L^* = \inf_{\hat{y}}{}^{a} \{R_L(\hat{y}) \mid \hat{y} \text{ measurable}^{b}\}.$$

A function $\hat{y}_L^*$ which minimizes the above functional is called **Bayes optimal learning rule** (with respect to the loss L).

[a] $\inf$ denotes the infimum (of a set)

[b] A measurable function is a function between the underlying sets of two measurable spaces. For example, in a classification problem (i.e., $\mathcal{Y}$ is a countable finite set) with input features in $\mathcal{X} = \Re^d$, $\hat{y}$ is a function between the measurable spaces $(\Re^d, \mathcal{B}(\Re^d))$ and $(\mathcal{Y}, 2^{\mathcal{Y}})$

Note: since we minimize over all measurable $\hat{y}$, the minimizer of $\mathbb{E}\big[L(\hat{y}(X), Y)\big]$ can be found by **pointwise minimization** of

$$\mathbb{E}[L(\hat{y}(X), Y) | X = x]$$

Bayes optimal risk – Examples

Classification:

0-1-loss: $L(\hat{y}(x), y) = \mathbb{1}_{\hat{y}(x) \neq y}$

$$\begin{aligned}\mathbb{E}[L(\hat{y}(X), Y) | X = x] &= \sum_{y \in \mathcal{Y}} L(\hat{y}(x), y)\, P(Y = y | X = x) \\ &= \sum_{y \in \mathcal{Y}} \mathbb{1}_{\hat{y}(x) \neq y}\, P(Y = y | X = x)\end{aligned}$$

Regression:

squared loss: $L(\hat{y}(x), y) = (y - \hat{y}(x))^2$

$$\begin{aligned}\mathbb{E}[L(\hat{y}(X), Y) | X = x] &= \int_{\mathcal{Y}} L(\hat{y}(x), y)\, p(y | X = x)\, dy \\ &= \int_{\mathcal{Y}} (y - \hat{y}(x))^2\, P(Y = y | X = x)\, dy.\end{aligned}$$

Outline

1 Bibliography

2 Introduction

3 Bayes rule

4 Bayes classifier

5 Cost-sensitive class.

6 Margin-based class.

7 Multi-class class.

8 Regression

Bayes classifier

Binary Classification: $\mathcal{Y} = \{-1, 1\}$.

0-1-**loss:** $L(\hat{y}(x), y) = \mathbb{1}_{\hat{y}(x)y \leq 0}$ is the canonical loss for classification.

In this case, the risk corresponds to the **probability of error**:

$$R(\hat{y}) = \mathbb{E}\big[\mathbb{1}_{\hat{y}(X)Y \leq 0}\big] = P(\hat{y}(X)Y \leq 0) = P(\hat{y}(X) \neq Y) = P(error).$$

Bayes classifier

Binary Classification: $\mathcal{Y} = \{-1, 1\}$.

0-1-**loss:** $L(\hat{y}(x), y) = \mathbb{1}_{\hat{y}(x)y \leq 0}$ is the canonical loss for classification.

In this case, the risk corresponds to the **probability of error**:

$$R(\hat{y}) = \mathbb{E}\big[\mathbb{1}_{\hat{y}(X)Y \leq 0}\big] = P(\hat{y}(X)Y \leq 0) = P(\hat{y}(X) \neq Y) = P(error).$$

Minimizaton of the risk: The risk (and thus probability of error) is minimized by the Bayesian decision rule since the risk decomposes as:

$$\begin{aligned} R(f) &= \mathbb{E}\big[\mathbb{1}_{\hat{y}(X)Y \leq 0}\big] = \mathbb{E}_X\big[\mathbb{E}_{Y|X}[\mathbb{1}_{\hat{y}(X)Y \leq 0}|X]\big] \\ &= \mathbb{E}_X[\mathbb{1}_{\hat{y}(X)=-1}P(Y=1|X) + \mathbb{1}_{\hat{y}(X)=1}P(Y=-1|X)]. \end{aligned}$$

Definition

The **Bayes classifier** $\hat{y}^* : \mathcal{X} \to \{-1, 1\}$ minimizing the error probablity is

$$\hat{y}^*(x) = \begin{cases} +1 & \text{if} \quad P(Y=1|X=x) > P(Y=-1|X=x) \\ -1 & \text{else} \end{cases}$$

Regression function

Definition

The **regression function**[a] $\eta(x)$ (of a binary classidication problem) is defined as

$$\eta(x) = \mathbb{E}[Y|X = x].$$

[a] Not to be confused with a regression problem.

Binary classification with $\mathcal{Y} = \{-1, 1\}$,[1]

$$\begin{aligned}\eta(x) &= \mathbb{E}[Y|X = x] = P(Y = 1|X = x) - P(Y = -1|X = x)\\ &= 2P(Y = 1|X = x) - 1.\end{aligned}$$

Bayes classifier as a margin-based classifier:

$$\hat{y}^*(x) = \operatorname{sign} \eta(x).$$

[1] Equivalently, if $\mathcal{Y} = \{0, 1\}$, then $\eta(x) = \mathbb{E}[Y|X = x] = P(Y = 1|X = x)$.

Bayes error

The **Bayes error** (risk of the Bayes classifier):

$$\begin{aligned} R^* &= \mathbb{E}_X\left[\min\{P(Y=1|X), P(Y=-1|X)\}\right] \\ &= \int_{\mathcal{X}} \min\{p(x|Y=1)P(Y=1), p(x|Y=-1)P(Y=-1)\}\, dx. \end{aligned}$$

$$\Longrightarrow \qquad 0 \leq R^* \leq \frac{1}{2}$$

Bayes error

The **Bayes error** (risk of the Bayes classifier):

$$\begin{aligned} R^* &= \mathbb{E}_X\left[\min\{P(Y=1|X), P(Y=-1|X)\}\right] \\ &= \int_{\mathcal{X}} \min\{p(x|Y=1)P(Y=1), p(x|Y=-1)P(Y=-1)\}\, dx. \end{aligned}$$

$$\implies \qquad 0 \leq R^* \leq \frac{1}{2}$$

Proposition

The Bayes risk R^* satisfies,

$$R^* \leq \min\{P(Y=1), P(Y=-1)\}.$$

To do: Proof.
Additional results: Error bounds for Normal features (Chapter 2.8 [DHS]).

Outline

1. Bibliography
2. Introduction
3. Bayes rule
4. Bayes classifier
5. Cost-sensitive class.
6. Margin-based class.
7. Multi-class class.
8. Regression

Cost-sensitive classification

Problem: Cost of errors is not always equal.

Example: Cancer detection from x-ray images
(cancer $Y = 1$, no cancer $Y = -1$)
cost of not detecting cancer (false negatives) is much higher than wrongly assigning a healthy person to be ill
(false positives).

	positive Prediction	negative Prediction
positive cases	true positives	false negatives
negative cases	false positives	true negatives

Cost matrix and Risk

Here, we refer to as $\hat{y}_c(X)$ to a cost-sensitive classifier, where different errors are penalized differently according to a cost matrix, i.e.,

Cost matrix:

$$C_{ij} = C(Y = i,\, \hat{y}_c(X) = j).$$

	positive Prediction	negative Prediction
positive cases	0	$C(Y=1,\, \hat{y}_c(X) = -1)$
negative cases	$C(Y=-1,\, \hat{y}_c(X) = 1)$	0

Cost sensitive 0-1-loss and risk:

$$\begin{aligned} R^C(f) &= \mathbb{E}\left[C(Y, \hat{y}_c(X))\, \mathbb{1}_{\hat{y}_c(X)Y \leq 0}\right] \\ &= \mathbb{E}_X[C_{1,-1}\, \mathbb{1}_{\hat{y}_c(X)=-1}\, P(Y=1|X) + C_{-1,1}\, \mathbb{1}_{\hat{y}_c(X)=1}\, P(Y=-1|X)]. \end{aligned}$$

Classification rule

Cost sensitive Bayes classifier:

$$\hat{y}_c^*(x) = \begin{cases} +1 & \text{if} \quad C_{1,-1}\, P(Y = 1|X = x) > C_{-1,1}\, P(Y = -1|X = x) \\ -1 & \text{else} \end{cases}$$

A new threshold for the regression function:

$$\hat{y}_c(x) = \operatorname{sign}\left[\eta(x) - \frac{C_{-1,1} - C_{1,-1}}{C_{-1,1} + C_{1,-1}}\right],$$

where $\eta(x) = \mathbb{E}[Y|X = x] = 2P(Y = 1|X = x) - 1$.

Observation: If $C_{-1,1} = C_{1,-1}$ (same costs for both classes), then we recover the standard Bayes classifier.
Homework: Derive the cost-sensitive Bayes classifier when $\mathcal{Y} = \{0, 1\}$.

Outline

1. Bibliography
2. Introduction
3. Bayes rule
4. Bayes classifier
5. Cost-sensitive class.
6. Margin-based class.
7. Multi-class class.
8. Regression

Margin-based classification

Classification Problem: We aim to learn a mapping function (classifier) of the form $\hat{y} : \mathcal{X} \to \{-1, 1\}$ that minimizes the 0-1-loss (and thus the probability of error). Unfortunately, finding a function that minimizes the 0-1-loss leads often to a hard optimization problem. Instead, we can minimize an alternative loss function which is easier to optimize.

Margin-based classification: Provides an "easier" approach to solve a classification problem as a regression problem by finding the function $f : \mathcal{X} \to \mathbb{R}$ that minimizes a surrogate convex loss, i.e., by :

- Using a **surrogate convex** loss function which upper bounds the 0-1-loss.
- Defining the classifier $\hat{y} : \mathcal{X} \to \{-1, 1\}$ as

$$\hat{y}(x) = \text{sign } f(x).$$

Note: In this lecture, we have assumed so far that we have access to the probability measure P on $\mathcal{X} \times \mathcal{Y}$. However, in practice, we only have access to *training data* $(x_i, y_i)_{i=1}^n$ sampled from the (unknown) probability measure P on $\mathcal{X} \times \mathcal{Y}$ (Lecture 4).

Loss function I

Definition (Convex margin-based loss function)

A function $L : \mathbb{R} \to \mathbb{R}_+$ is a **convex margin-based loss function** if

- $L(y, f(x)) = L(y\, f(x))$, where the function (of the product) $y\, f(x) \in \mathbb{R}$ is called the **functional margin**,
- L is convex[a],
- L upper bounds the 0-1-loss.

[a] see lecture 7 for formal definition of a convex function

Loss function I

Definition (Convex margin-based loss function)

A function $L : \mathbb{R} \to \mathbb{R}_+$ is a **convex margin-based loss function** if

- $L(y, f(x)) = L(y\, f(x))$, where the function (of the product) $y\, f(x) \in \mathbb{R}$ is called the **functional margin**,
- L is convex[a],
- L upper bounds the 0-1-loss.

[a] see lecture 7 for formal definition of a convex function

Examples:

hinge loss (soft margin loss)	$L(y\, f(x)) = \max(0, 1 - y\, f(x))$
truncated squared loss	$L(y\, f(x)) = \max(0, 1 - y\, f(x))^2$
exponential loss	$L(y\, f(x)) = \exp(-y\, f(x))$
logistic loss	$L(y\, f(x)) = \log(1 + \exp(-y\, f(x)))$

Loss function II

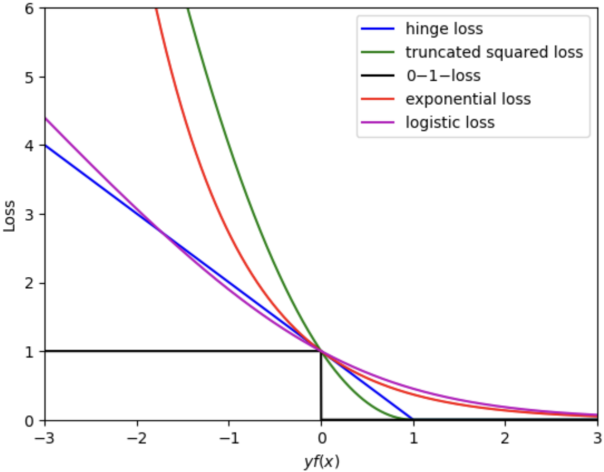


Figure: Plots for different classification loss functions

Optimality I

Problem: Different loss measures $\Longrightarrow$ Different optimal function

Key question: Let, $f_L^* : \mathcal{X} \to \mathbb{R}$, be the function which minimizes the risk R_L,

$$R_L(f) = \mathbb{E}\big[L(f(X)Y)\big],$$

where L is a convex margin-based loss function (surrogate of the 0-1-loss). **Does the sign of f_L^* agree with the Bayes classifier $\hat{y}^*(x)$?** I.e.,

$$\hat{y}^*(x) \overset{?}{=} \operatorname{sign} f_L^*(x).$$

Optimality I

Problem: Different loss measures $\Longrightarrow$ Different optimal function

Key question: Let, $f_L^* : \mathcal{X} \to \mathbb{R}$, be the function which minimizes the risk R_L,

$$R_L(f) = \mathbb{E}\big[L(f(X)Y)\big],$$

where L is a convex margin-based loss function (surrogate of the 0-1-loss). **Does the sign of f_L^* agree with the Bayes classifier $\hat{y}^*(x)$?** I.e.,

$$\hat{y}^*(x) \overset{?}{=} \text{sign } f_L^*(x).$$

Definition

A margin-based loss function $L : \mathbb{R} \to [0, \infty)$ is **classification calibrated** if for all $\eta(x) \neq 0$, then

$$\text{sign } f_L^*(x) = \hat{y}^*(x) = \text{sign}\, \eta(x),$$

i.e., f_L^* has the same sign as the Bayes classifier $\hat{y}^*$.

Reminder:

$\eta(x) = \mathbb{E}[Y|X = x] = P(Y = 1|X = x) - P(Y = -1|X = x)$

Optimality II

Cost sensitive risk functional based on convex margin-based loss:

$$R_L^C(f) = \mathbb{E}_X[C_{1,-1}\, L(f(X))\, P(Y=1|X) + C_{-1,1}\, L(-f(X))\, P(Y=-1|X)]$$

$$f^*_{C,L} = \arg\min\{R_L^C(f)\,|\, f \text{ measurable}\}.$$

Definition

A margin-based loss function $L : \mathbb{R} \to [0,\infty)$ is **cost-sensitive classification calibrated** if for all $\eta(x) \neq \frac{C_{-1,1}-C_{1,-1}}{C_{1,-1}+C_{-1,1}}$ we have

$$\operatorname{sign} f^*_{C,L}(x) = \hat{y}^*_C(x) = \operatorname{sign}\left[\eta(x) - \frac{C_{-1,1}-C_{1,-1}}{C_{1,-1}+C_{-1,1}}\right],$$

that is $f^*_{C,L}$ has the same sign as the cost-senitive Bayes classifier $\hat{y}^*_C$.

Optimality III

Loss	Loss function $L(y\,f(x))$	Optimal function
hinge (soft-margin)	$\max(0, 1 - y\,f(x))$	$f_L^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 0 \\ -1 & \text{if } \eta(x) < 0 \end{cases}$
truncated squared	$\max(0, 1 - y\,f(x))^2$	$f_L^*(x) = \eta(x),$
exponential	$\exp(-y\,f(x))$	$f_L^*(x) = \frac{1}{2} \log \frac{1+\eta(x)}{1-\eta(x)},$
logistic	$\log(1 + \exp(-y\,f(x)))$	$f_L^*(x) = \log \frac{1+\eta(x)}{1-\eta(x)}.$

The loss functions together with their minimizers $f_L^*(x)$ in terms of the regression function $\eta(x) = \mathbb{E}[Y|X = x] = P(Y = 1|X = x) - P(Y = -1|X = x)$.

Homework: Which of the above margin-based loss functions are classification calibrated?

Outline

1 Bibliography

2 Introduction

3 Bayes rule

4 Bayes classifier

5 Cost-sensitive class.

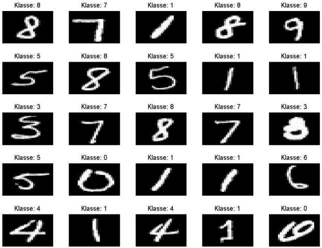
6 Margin-based class.

7 Multi-class class.

8 Regression

Multi-class Classification

$\mathcal{Y} = \{1, \dots, K\}$ (no order!)



Multi-class risk of the 0-1-loss:

$$R(\hat{y}) = \mathbb{E}\big[\mathbb{1}_{\hat{y}(X)\neq Y}\big] = \mathbb{E}\big[\,\mathbb{E}[\mathbb{1}_{\hat{y}(X)\neq Y}|X]\,\big] = \mathbb{E}\Big[\sum_{k=1}^{K} \mathbb{1}_{\hat{y}(X)\neq k} P(Y=k|X)\Big].$$

Multi-class Bayes classifier:

$$\hat{y}^*(x) = \operatorname*{arg\,max}_{k\in\{1,\dots,K\}} P(Y=k|X=x),$$

Multi-class Bayes risk:

$$R^* = \mathbb{E}\Big[1 - \max_{k\in\{1,\dots,K\}} P(Y=k|X)\Big].$$

Multi-class Classification II

Idea: Decompose multi-class problem into binary classification problems,

- **one-vs-all**: The multi-class problem is decomposed into K binary problems. Each class versus all other classes ⇒ K classifiers $\{f_i\}_{i=1}^K$.

  $$f_{OVA}(x) = \arg\max_{i=1,\ldots,K} f_i(x),$$

  where ideally $f_i(x) = P(Y = i|x)$.

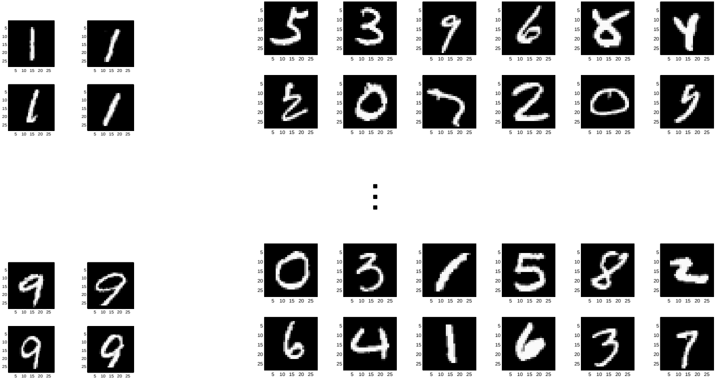
- **one-vs-one**: The multi-class problem is decomposed into $\binom{K}{2}$ binary problems. Each class versus each other class. Each binary classifier f_{ij} votes for one class. Final classification by majority vote,

  $$f_{OVO}(x) = \arg\max_{i=1,\ldots,K} \sum_{\substack{j=1 \\ j\neq i}}^{K} \mathbb{1}_{f_{ij}(x)>0},$$

  where ideally $f_{ij}(x) = P(Y = i|x) - P(Y = j|x)$.

One-vs-all

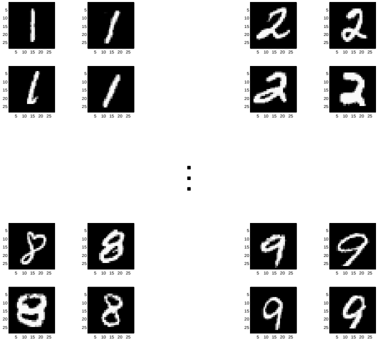
Decompose multi-class problem into K binary classification problems,



Handwritten digits: $K = 10 \implies$ 10 binary classification problems.

One-vs-one

Decompose multi-class problem into $\binom{K}{2}$ binary classification problems,



Handwritten digits: $K = 10 \implies$ 45 binary classification problems.

Optimality

Theorem

The one-vs-all and one-vs-one multi-class schemes lead to the Bayes optimal solution for the multi-class problem if the binary classifiers f_i and f_{ij} for all $i, j \in \mathcal{Y}$ are **strictly monotonically increasing functions of the conditional distribution***.*

Proof.

One-vs-all: Given that f_i are strictly monotonically increasing functions of the conditional distribution, i.e., $f_i(x) = g(P(Y = i|X = x))$ with $g(\cdot)$ being a strictly monotonically increasing function, we have that

$$\arg\max_{i=1,\dots,K} f_i(x) = \arg\max_{i=1,\dots,K} g(P(Y = i|X = x)) = \arg\max_{i=1,\dots,K} P(Y = i|X = x) = \hat{y}^*.$$

Optimality

Theorem

The one-vs-all and one-vs-one multi-class schemes lead to the Bayes optimal solution for the multi-class problem if the binary classifiers f_i and f_{ij} for all $i, j \in \mathcal{Y}$ are **strictly monotonically increasing functions of the conditional distribution***.*

Proof. *One-vs-one:* Given that f_{ij} are strictly monotonically increasing functions of the conditional dstribution, i.e., $f_{ij}(x) = g(P_{ij}(Y = i|x))$ with $P_{ij}(Y = i|x) = \frac{P(Y=i|X=x)}{P(Y=i|X=x)+P(Y=j|X=x)}$, and that the binary optimal classifier fulfills that $f^*_{ij} = -f^*_{ji}$, then

$$\arg\max_{i=1,\ldots,K} \sum_{\substack{j=1\\j\neq i}}^{K} \mathbb{1}_{f^*_{ij}(x)>0} = \arg\max_{i=1,\ldots,K} \sum_{\substack{j=1\\j\neq i}}^{K} \mathbb{1}_{f^*_{ij}(x)>f^*_{ji}(x)} = \arg\max_{i=1,\ldots,K} \sum_{\substack{j=1\\j\neq i}}^{K} \mathbb{1}_{g(P_{ij}(Y=i|x))>g(P_{ij}(Y=j|x))}$$

$$= \arg\max_{i=1,\ldots,K} \sum_{\substack{j=1\\j\neq i}}^{K} \mathbb{1}_{P_{ij}(Y=i|x)>P_{ij}(Y=j|x)} = \arg\max_{i=1,\ldots,K} \sum_{\substack{j=1\\j\neq i}}^{K} \mathbb{1}_{P(Y=i|x)>P(Y=j|x)} = \arg\max_{i=1,\ldots,K} P(Y = i|x)$$

Outline

1. Bibliography
2. Introduction
3. Bayes rule
4. Bayes classifier
5. Cost-sensitive class.
6. Margin-based class.
7. Multi-class class.
8. Regression

Regression

Regression: output space $\mathcal{Y} = \mathbb{R}$,
Risk: $R(f) = \mathbb{E}\big[L(Y, f(X))\big] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[L(Y, f(X) \,|\, X]\right.$
Loss function: $L(y, f(x))$.

Loss function	Optimal regressor
Squared loss: $L(y, f(x)) = (y - f(x))^2$	$f^*_L(x) = \mathbb{E}_Y[Y \vert X = x]$
L_1 - **loss:** $L(y, f(x)) = \vert y - f(x) \vert$	$f^*_L(x) = \text{Median}(Y \vert X = x)$
ε**-insensitive :** $L(y, f(x)) = (\vert y - f(x)\vert - \varepsilon)\mathbb{1}_{\vert y - f(x)\vert > \varepsilon}$	not unique
Huber's robust loss: $L(y, f(x)) = \begin{cases} \frac{1}{2\epsilon}(y - f(x))^2 & \text{if } \vert y - f(x)\vert \leq \varepsilon \\ \vert y - f(x)\vert - \frac{\varepsilon}{2} & \text{if } \vert y - f(x)\vert > \varepsilon \end{cases}$	unknown

Observation: In regression problems, the optimal regression function depends on the considered loss. I.e., there is not an equivalent to the Bayes classifier for regression problems.

Loss functions for regression

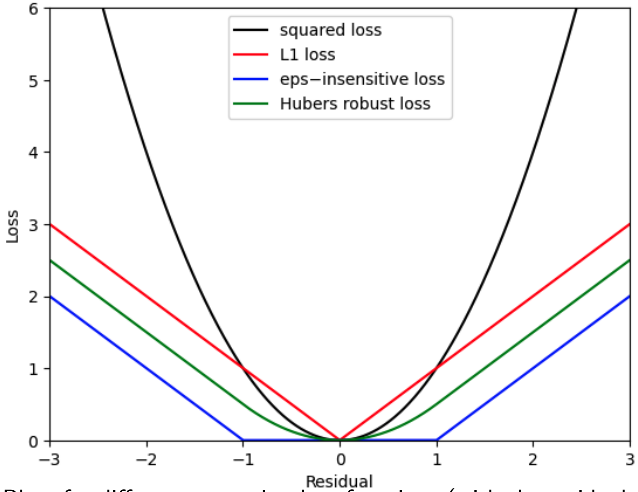


Figure: Plots for different regression loss functions (with the residual $y - f(x)$ as argument)

Median is more stable than the mean

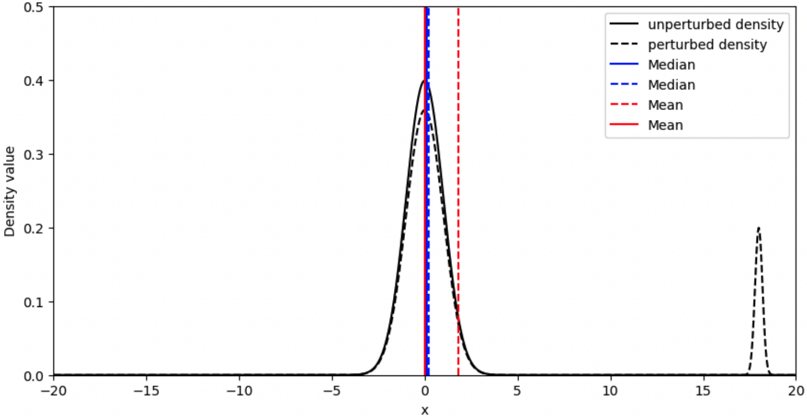


Figure: Robustness of the median versus the mean. Here, we observe that when the underlying feature density is affected by perturbations (e.g., outliers), the mean value is significantly shifted. In contrast, the media remains almost unperturbed.

Outline

Summary

- Bayesian decision theory allows us to make optimal decisions under uncertainty.
- The optimal binary classifier (in terms of error probability) is the Bayes classifier and selects the class that maximizes the posterior $P(Y|x)$ for each feature vector x.
- Bayes classifier can be extended to *cost-sensitive learning* and the *multi-class* setting. For multi-class problems we have seen two approaches: one-versus-all and one-versus-one.
- Margin-based classifiers allows us to solve classification problems by minimizing a surrogate loss function that is easier to optimize than the 0-1-loss.
- In contrast, in regression problems, the optimal regression function is loss-dependent (as we do not have a canonical choice for measuring errors).
- Next lecture we will see how to optimally solve regression and classification problems using data!