Lecture 3: Bayesian Decision Theory

Prof. Dr. Mario Fritz

2022 04 25

https://fritz.cispa.saarland
Trustworthy AI
CISPA Helmholtz Center for Information Security, Saarbrücken, Germany

Outline

Main references

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

Outline

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 $\mathrm{P}(y|x)$ that determines the probability of observing class y for the given features 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?

Notation

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

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

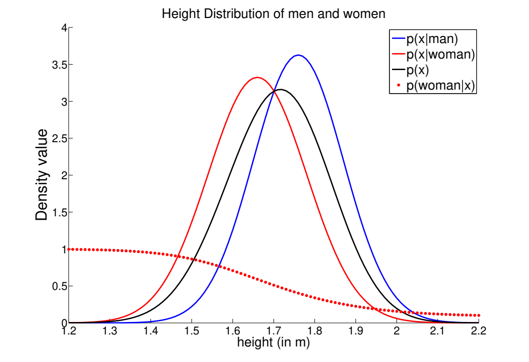
where

- $\mathrm{P}(y|x)$ denotes the **posterior probability** and corresponds to the probability that we observe y after observing x.
- $p(x|y)$ denotes the **class-conditional density (or likelihood)** and models the occurrence of the features x of class y.
- $\mathrm{P}(y)$ 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.
- $\mathrm{P}(x)$ denotes the **marginal distribution (or evidence)** of the features x and models the cumulated occurrence of features over all classes $y \in \mathcal{Y}$.

Example I

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

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



Example II

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

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

$\rightarrow$ Always decides same class for all x. $P(error|x) = P(error) = \min[Pr(male), \mathrm{P}(female)]$.

$\rightarrow$ For an observed feature vector x, $P(error|x) = \min[Pr(x|male), \mathrm{P}(x|female)]$.

$\rightarrow$ For an observed feature vector x, $P(error|x) = \min[Pr(male|x), \mathrm{P}(female|x)]$.

Example II

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

$$p(x/\omega_j) \qquad P(\omega_1) = \frac{2}{3} \quad P(\omega_2) = \frac{1}{3} \qquad P(\omega_j/x)$$

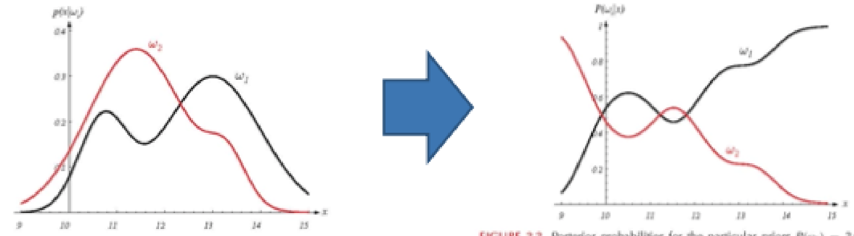


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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.

Bayes Decision Rule

The **Bayes (optimal) decision rule** given by:

$$y^* = \arg\max_i \mathrm{P}(\omega_i|x),$$

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

$$P(error|x) = \min[Pr(\omega_1|x), \mathrm{P}(\omega_2|x).]$$

and

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

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

Loss function and risk

Quantitative measure of error:

Definition (Loss function)

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

Examples:

Classification:	0-1-loss,	$L(\hat{y}(x), y) = \mathbb{1}_{\hat{y}(x) \neq y}$
Regression:	squared loss,	$L(\hat{y}(x), y) = (y - \hat{y}(x))^2$

Loss function and risk

Quantitative measure of error:

Definition (Loss function)

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

Examples:

Classification:	0-1-loss,	$L(\hat{y}(x), y) = \mathbb{1}_{\hat{y}(x) \neq y}$
Regression:	squared loss,	$L(\hat{y}(x), y) = (y - \hat{y}(x))^2$

Definition (Risk)

The **risk** or **expected loss** of a learning rule $f : \mathcal{X} \to \mathcal{Y}$ 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: $\mathbb{E}\Big[\mathbb{E}[L(\hat{y}(X), Y)|X]\Big] \; = \int_{\mathbb{R}^d} \Big[\int_{\mathbb{R}} L(\hat{y}(x), y)\, p(y|x) dy \Big] \, p(x) \, dx.$

Bayes optimal risk

Definition

The **Bayes optimal risk** is given by

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

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

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

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

Classification: $\mathbb{E}[L(\hat{y}(X), Y)|X = x] = \sum_{y \in \mathcal{Y}} L(\hat{y}(x), y)\, \mathrm{P}(Y = y|X = x).$

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

Outline

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.

Risk is the **probability of error**:

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

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.

Risk is the **probability of error**:

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

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

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

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

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

Regression function

Definition

The **regression function** $\eta(x)$ is defined as

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

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

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

Bayes classifier as a margin-bassed classifier:

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

Bayes error

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

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

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

Bayes error

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

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

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

Proposition

The Bayes risk R^* satisfies,

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

To do: Proof.

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

Outline

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:

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

	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:**

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

Classification rule

Cost sensitive Bayes classifier:

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

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

Observation : If $C_{-1,1} = C_{1,-1}$ (same costs for both classes), then we recover the standard Bayes classifier.

Outline

Margin-based classification

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).

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 fucntion 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

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

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

Loss function I

Definition (Convex margin-based loss function)

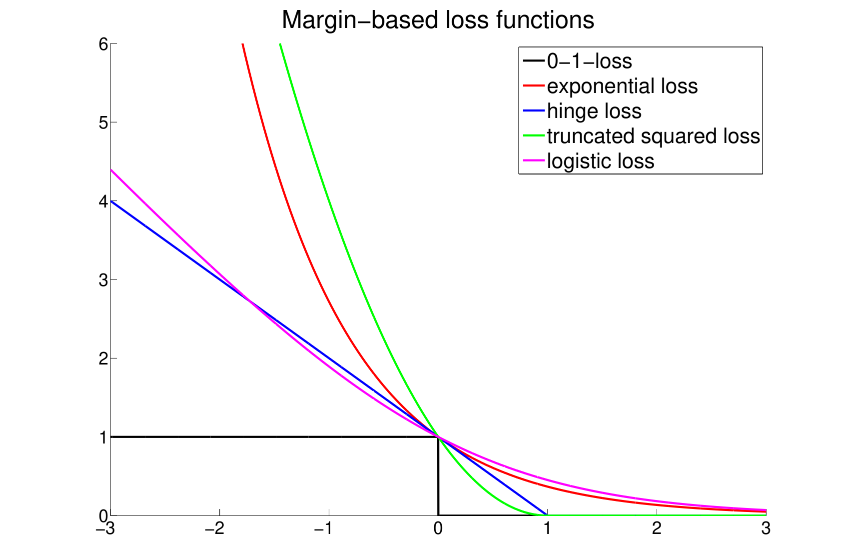
A function $L : \mathbb{R} \to \mathbb{R}_+$ is a **convex margin-based loss function** if

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

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



Optimality I

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

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.,

$$\hat{y}^*(x) \stackrel{?}{=} \mathrm{sign}\, f_L^*(x).$$

Optimality I

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

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.,

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

Definition

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

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

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

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

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

$$R_L^C(f) = \mathbb{E}_X[C_{1,-1}\, L(f(X))\, \mathrm{P}(Y=1|X) + C_{-1,1}\, L(-f(X))\, \mathrm{P}(Y=-1|X)]$$
$$f^*_{C,L} = \text{argmin}\, \{R_L^C(f) \,|\, f \text{ measurable}\}.$$

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

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

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

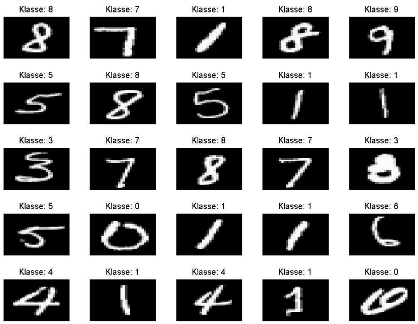
Optimality III

Examples of surrogate convex losses for classification with their optimal solution:

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	$\log(1 + \exp(-y\,f(x)))$	$f_L^*(x) = \log\frac{1+\eta(x)}{1-\eta(x)}$.

Outline

Multi-class Classification



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

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

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

Multi-class Bayes classifier:

$$\hat{y}^*(x) = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}} \mathrm{P}(Y = k|X = x),$$

Multi-class Bayes risk:

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 $\Rightarrow$ K classifiers $\{f_i\}_{i=1}^K$.

$$f_{OVA}(x) = \operatorname*{argmax}_{i=1,\dots,K} f_i(x),$$

  where ideally $f_i(x) = \mathrm{P}(Y = i|x)$.

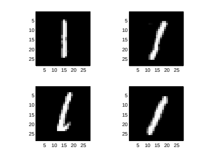
- **one-vs-one**: The multi-class problem is decomposed into $\binom{K}{2}$ binary problems. Each class versus each other class. Each binary classifier f_{ij} votes for one class. Final classification by majority vote,

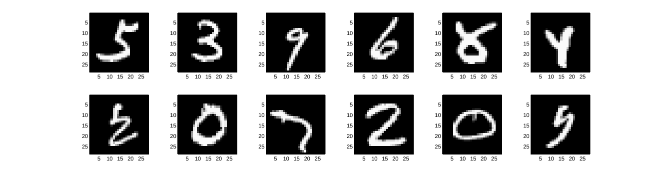
$$f_{OVO}(x) = \operatorname*{argmax}_{i=1,\dots,K} \sum_{\substack{j=1 \\ j \neq i}}^{K} \mathbb{1}_{f_{ij}(x)>0},$$

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

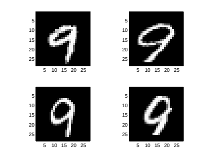
One-vs-all

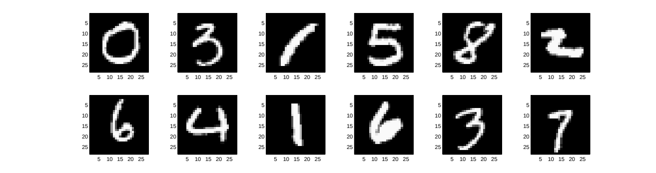
Decompose multi-class problem into K binary classification problems,





$\vdots$

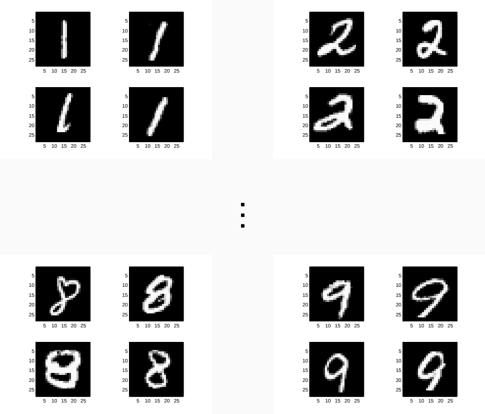




Handwritten digits: $K = 10 \implies 10$ binary classification problems.

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., $f_{ij}(x) = g(\mathrm{P}(Y = i|X = x))$ with $g()$ being a strictly monotonically increasing function, we have that

$$\operatorname*{argmax}_{i=1,\dots,K} f_i(x) = \operatorname*{argmax}_{i=1,\dots,K} g(\mathrm{P}(Y = i|X = x)) = \operatorname*{argmax}_{i=1,\dots,K} Pr(Y = i|X = x) = \hat{y}^*.$$

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., $f_{ij}(x) = g(\mathrm{P}_{ij}(Y = i|x))$ with $\mathrm{P}_{ij}(Y = i|x) = \frac{\mathrm{P}(Y=i|X=x)}{\mathrm{P}(Y=i|X=x)+\mathrm{P}(Y=j|X=x)}$, and that the binary optimal classifier fulfills that $f^*_{ij} = -f^*_{ji}$, then

$$\operatorname*{argmax}_{i=1,\dots,K} \sum_{\substack{j=1\\j\neq i}}^{K} \mathbb{1}_{f^*_{ij}(x)>0} = \operatorname*{argmax}_{i=1,\dots,K} \sum_{\substack{j=1\\j\neq i}}^{K} \mathbb{1}_{f^*_{ij}(x)>f^*_{ji}(x)} = \operatorname*{argmax}_{i=1,\dots,K} \sum_{\substack{j=1\\j\neq i}}^{K} \mathbb{1}_{g(\mathrm{P}_{ij}(Y=i|x))>g(\mathrm{P}_{ij}(Y=j|x))}$$

$$= \operatorname*{argmax}_{i=1,\dots,K} \sum_{\substack{j=1\\j\neq i}}^{K} \mathbb{1}_{\mathrm{P}_{ij}(Y=i|x)>\mathrm{P}_{ij}(Y=j|x)} = \operatorname*{argmax}_{i=1,\dots,K} \sum_{\substack{j=1\\j\neq i}}^{K} \mathbb{1}_{\mathrm{P}(Y=i|x)>\mathrm{P}(Y=j|x)} = \operatorname*{argmax}_{i=1,\dots,K} \mathrm{P}(Y = i|x)$$

Outline

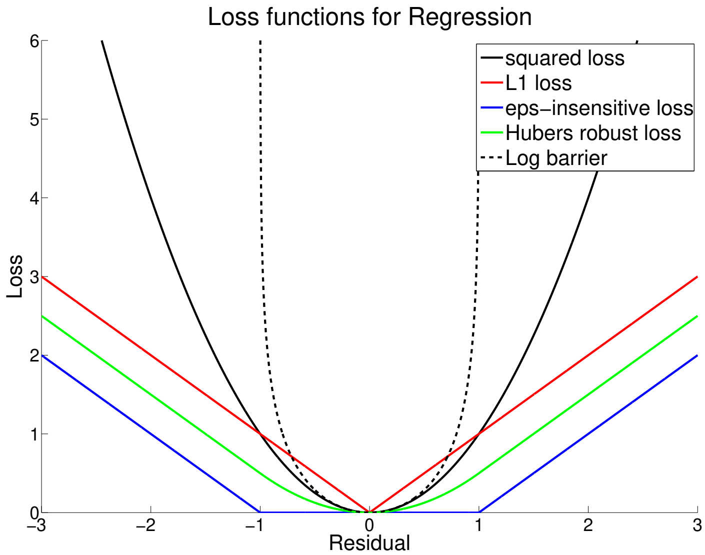
Regression

Regression: output space $\mathcal{Y} = \mathbb{R}$,
Risk: $R(f) = \mathbb{E}\big[L(Y, f(X))\big] = \mathbb{E}_X\big[\mathbb{E}_{Y|X}[L(Y, f(X) \mid X]$
Loss function: $L(y, f(x))$ (often plotted with $|y - f(x)|$ as argument).

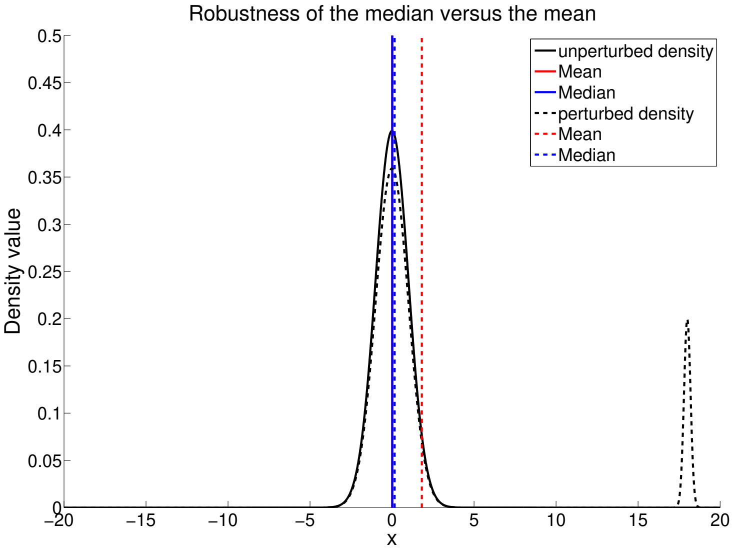
Loss function	Optimal regressor
Squared loss: $L(y, f(x)) = (y - f(x))^2$	$f_L^*(x) = \mathbb{E}_Y[Y \mid X = x]$
L_1 **- loss:** $L(y, f(x)) = \lvert y - f(x)\rvert$	$f_L^*(x) = \mathrm{Median}(Y \mid X = x)$
ε**-insensitive :** $L(y, f(x)) = (\lvert y - f(x)\rvert - \varepsilon)\mathbb{1}_{\lvert y-f(x)\rvert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

Observation: In regression problems, the optimal regression function depends on the loss.

Loss functions for regression III



Median is more stable than the mean



Outline

Summary

- Bayesian decision theory allows us to make optimal decisions under uncertainty.
- The optimal binary classifier 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.
- Next lecture we will see how to solve regression and classification problems using data!