Machine Learning

Bayesian Decision Theory

Prof. Matthias Hein

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

Lecture 3, 25.10.2013

Statistical learning I

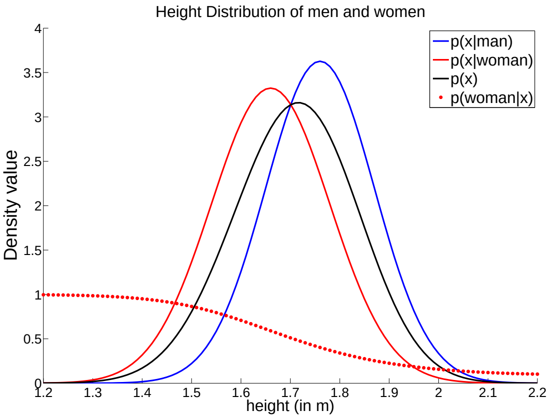
- Assumption: Data is generated by a **probability measure** P on $\mathcal{X} \times \mathcal{Y}$.
- What does that mean ?
  1. Training data is a **random sample** from P,
  2. The labels $y \in \mathcal{Y}$ are **non-deterministic**, that means there exists not necessarily a function $y = g(x)$. Instead for a given feature x, there exists a distribution over the possible values in $\mathcal{Y}$.
  3. Since the training data underlies statistical fluctuations, the classifier should be relatively stable under small changes of the training data.

Setting: binary classification, that is $\mathcal{Y} = \{-1, 1\}$.
The **joint density** $p(x, y)$ of the probability measure P on $\mathcal{X} \times \mathcal{Y}$ can be decomposed as follows

- The **class-conditional density** $p(x|y)$. It models the occurrence of the features x of class y.
- The **conditional probability** $\mathrm{P}(y|x)$. The probability that we observe y given that the input is x. The most probable class y for the features x is then used for prediction.
- The **marginal distribution** $p(x)$. It models the cumulated occurrence of features x over all classes.
- The **class probabilities** $\mathrm{P}(y)$. The total probability of a class y.

Learning problem: Predict sex of a person using height as feature.

- input space $\mathcal{X} = \mathbb{R}$,
- output space: $Y = \{\mathrm{male}, \mathrm{female}\}$.



Marginal distribution

$$p(x) = p(x|\mathrm{male})\mathrm{P}(\mathrm{male}) + p(x|\mathrm{female})\mathrm{P}(\mathrm{female}).$$

Using Bayes law we get the conditional probability $\mathrm{P}(y|x)$,

$$\mathrm{P}(y|x) = \frac{p(x|y)\mathrm{P}(y)}{p(x)}.$$

Classification rule: classify x as female if $\mathrm{P}(\mathrm{female}|x) \geq \frac{1}{2}$ and otherwise as male.

$\Longrightarrow$ From the plot, female if $x < 1.71$ and otherwise male.

Generally there is **no** deterministic relation $Y = g(X)$!

but !

Probability distribution over the possible values $\mathrm{P}(y|x)$

Bayesian decision theory:
What is the optimal classifier/function given a way how to measure the difference between the output $f(X)$ and Y ?

or

How to make optimal decisions under uncertainty ?

Loss function and risk

Quantitative measure of error:

Definition

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

Examples:

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

Loss function and risk

Quantitative measure of error:

Definition

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

Examples:

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

Definition

The **risk** or **expected loss** of a learning rule f is defined as

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

How to interpret $\mathbb{E}\Big[\mathbb{E}[L(f(X), Y)|X]\Big]$ (here: $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$):

$$\mathbb{E}\Big[\mathbb{E}[L(f(X), Y)|X]\Big] = \int_{\mathbb{R}^d} \Big[\int_{\mathbb{R}} L(f(x), y)\, p(y|x) dy \Big]\, p(x)\, dx.$$

Bayes optimal risk

Definition

The **Bayes optimal risk** is given by

$$R_L^* = \inf_f \{R(f) \mid f \text{ measurable}\}.$$

A function f_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 f, the minimizer of $\mathbb{E}\, L(f(X), Y)$ can be found by **pointwise minimization** of

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

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

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

Bayes classifier

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

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

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

Risk is the **probability of error** !

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

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

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

Risk is the **probability of error** !

Decomposition of the risk:

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

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

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

Bayes classifier II

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:

$$f^*(x) = \mathrm{sign}\, \eta(x).$$

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):

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

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

Proposition

The Bayes risk R^* satisfies,

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

and for any measurable mapping $\phi : \mathcal{X} \to \mathcal{Z}$ we have

$$R^*_{\mathcal{X}} \leq R^*_{\mathcal{Z}}.$$

Bayes error II

- **Example:** $\mathrm{P}(Y = 1) = 0.95$ and $\mathrm{P}(Y = -1) = 0.05$,

$$R^* \leq \min\{\mathrm{P}(Y = 1), \mathrm{P}(Y = -1)\} = 0.05.$$

The upper bound can always be achieved. Take

$$f(x) = \begin{cases} 1 & \text{if } \mathrm{P}(Y = 1) > \mathrm{P}(Y = -1) \\ -1 & \text{else} \end{cases}.$$

$\Rightarrow$ Learning is difficult if classes are heavily disbalanced.

- Transformations of the data can never decrease the error.

Example: $\mathcal{X} = \mathbb{R}$, $\left.\begin{array}{c} \mathrm{P}(Y = 1 | X = x) = 1 \text{ if } x < 0 \\ \mathrm{P}(Y = -1 | X = x) = 1 \text{ if } x > 0 \end{array}\right\} \Longrightarrow R^*_{\mathcal{X}} = 0.$

Marginal distribution of X is symmetric around origin ($p(x) = p(-x)$).

Transformation: $Z = X^2$, $\mathrm{P}(Y = 1 | Z = z) = \frac{1}{2} \Longrightarrow R^*_Z = \frac{1}{2}$.

Decision boundary demo !

Problem: Minimization of 0-1-loss leads often to NP-hard problems

Solution:

- One uses convex surrogates which upper bound the 0-1-loss.
- The output space $\mathcal{Y} = \{-1, 1\}$ is relaxed to $\mathcal{Y} = \mathbb{R}$.
- Solve regression problem $g : \mathcal{X} \to \mathbb{R}$.
- Do classification with $f : \mathcal{X} \to \{-1, 1\}$, given by

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

Convex-margin based loss functions II

Definition

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

- $L(y, f(x)) = L(y\, f(x))$, $\quad$ $y\, f(x)$ is called the **functional margin**,
- L is convex,
- L upper bounds the 0-1-loss

$$\mathbb{1}_{\alpha \leq 0} \;\leq\; L(\alpha), \quad \forall\, \alpha \in \mathbb{R}.$$

Convex-margin based loss functions II

Definition

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

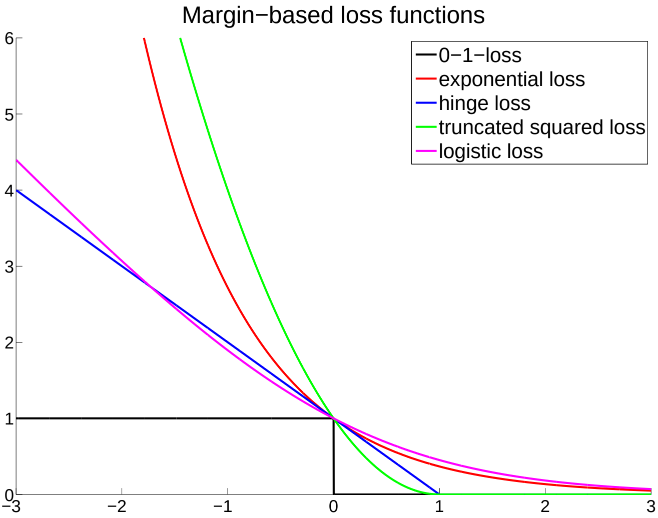
- $L(y, f(x)) = L(y\,f(x))$, $\quad$ $y\,f(x)$ is called the **functional margin**,
- L is convex,
- L upper bounds the 0-1-loss

$$\mathbb{1}_{\alpha \leq 0} \;\leq\; L(\alpha), \quad \forall\, \alpha \in \mathbb{R}.$$

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	$L(y\,f(x)) = \log_2(1 + \exp(-y\,f(x)))$

Convex margin-based loss functions III



Convex margin-based loss functions IV

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 ?

$$\text{Bayes classifier } f^*(x) \stackrel{?}{=} \mathrm{sign}\, f_L^*(x).$$

Convex margin-based loss functions IV

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 ?

$$\text{Bayes classifier } f^*(x) \stackrel{?}{=} \operatorname{sign} f_L^*(x).$$

Definition

A margin-based loss function $L : \mathbb{R} \to [0, \infty)$ is **classification calibrated** if for all $\eta(x) = \mathbb{E}[Y|X = x] \neq 0$ we have

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

that is f_L^* has the same sign as the Bayes classifier f^*.

Convex-margin based loss functions V

Theorem

Let L be a margin-based, convex loss function. Then L is **classification calibrated** *if and only if*

$$L \text{ is } \textbf{differentiable at } 0 \text{ and } \left.\frac{\partial L}{\partial x}\right|_{x=0} < 0.$$

$\Rightarrow$ Other loss functions are also classification calibrated e.g. squared loss.

Theorem

Let L be a margin-based, convex loss function. Then L is **classification calibrated** *if and only if*

$$L \text{ is } \textbf{differentiable at } 0 \text{ and } \left.\frac{\partial L}{\partial x}\right|_{x=0} < 0.$$

$\Rightarrow$ Other loss functions are also classification calibrated e.g. squared loss.

hinge loss	$L(y\,f(x)) = \max(0, 1 - y\,f(x))$	$f_L^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 0 \\ -1 & \text{if } \eta(x) < 0 \end{cases}$
tr. sqr. l.	$L(y\,f(x)) = \max(0, 1 - y\,f(x))^2$	$f_L^*(x) = \eta(x)$,
exp. loss	$L(y\,f(x)) = \exp(-y\,f(x))$	$f_L^*(x) = \frac{1}{2}\log\frac{1+\eta(x)}{1-\eta(x)}$,
log. loss	$L(y\,f(x)) = \log_2(1 + \exp(-y\,f(x)))$	$f_L^*(x) = \log\frac{1+\eta(x)}{1-\eta(x)}$.

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

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:

$$C_{ij} = C(Y = i, \mathrm{sign}(f(X)) = j).$$

	positive Prediction	negative Prediction
positive cases	0	$C(Y = 1, \mathrm{sign}(f(X)) = -1)$
negative cases	$C(Y = -1, \mathrm{sign}(f(X)) = 1)$	0

Cost sensitive 0-1-loss:

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

Cost sensitive Bayes classifier:

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

A new threshold for the regression function:

$$f_C^*(x) = \operatorname{sign}\Big[\eta(x) - \frac{C_{-1,1} - C_{1,-1}}{C_{1,-1} + C_{-1,1}}\Big],$$

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

If $C_{-1,1} = C_{1,-1}$ (same costs for both classes) $\Longrightarrow$ threshold is zero.

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

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

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

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

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

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

Cost-sensitive classification IV

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

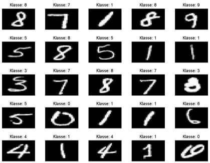
$$\operatorname{sign} f^*_{C,L}(x) = f^*_C(x) = \operatorname{sign}\left[\eta(x) - \frac{C_{-1,1} - C_{1,-1}}{C_{1,-1} + C_{-1,1}}\right],$$

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

Theorem

Let L be a convex margin-based loss function. Then L is **cost-sensitive classification calibrated** *if and only if*

$$L \text{ is differentiable at } 0 \text{ and } \left.\frac{\partial L}{\partial x}\right|_{x=0} < 0.$$



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

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

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

Multi-class Bayes classifier:

$$f^*(x) = \arg\max_{k\in\{1,\ldots,K\}} \mathrm{P}(Y=k|X=x),$$

Multi-class Bayes risk:

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

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 $\Rightarrow$ K classifiers $\{f_l\}_{l=1}^K$.

$$f_{OVA}(x) = \underset{l=1,\ldots,K}{\arg\max}\, f_l(x).$$

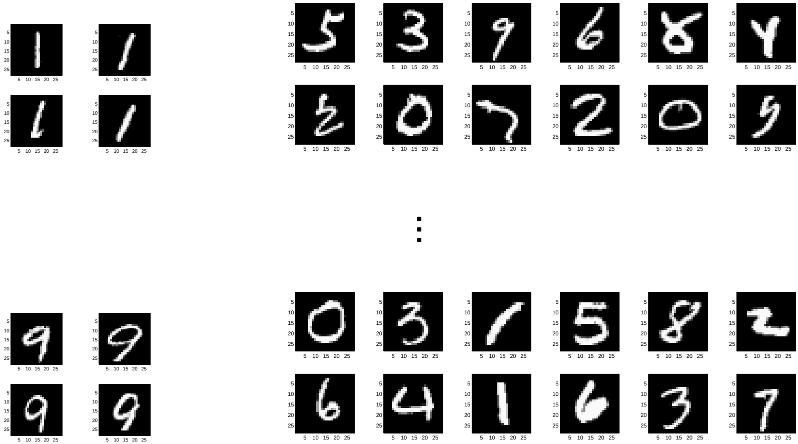
- **one-vs-one**: The multi-class problem is decomposed into $\binom{K}{2}$ binary problems. Each class versus each other class. Each binary classifier f_{lm} votes for one class. Final classification by majority vote,

$$f_{OVO}(x) = \underset{l=1,\ldots,K}{\arg\max} \sum_{\substack{m=1\\ m\neq l}}^{K} \mathbb{1}_{f_{lm}(x)>0}.$$

Multi-class Classification III

one-vs-all:

Decompose multi-class problem into K binary classification problems,

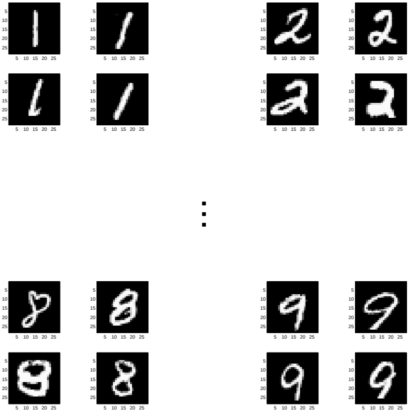


Handwritten digits: $K = 10 \Longrightarrow 10$ binary classification problems.

Multi-class Classification III

one-vs-one:

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



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

Multi-class Classification IV

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_l are **strictly monotonically increasing functions of the conditional distribution***.*

Loss functions for regression

Regression: output space $\mathcal{Y} = \mathbb{R}$,
Risk:

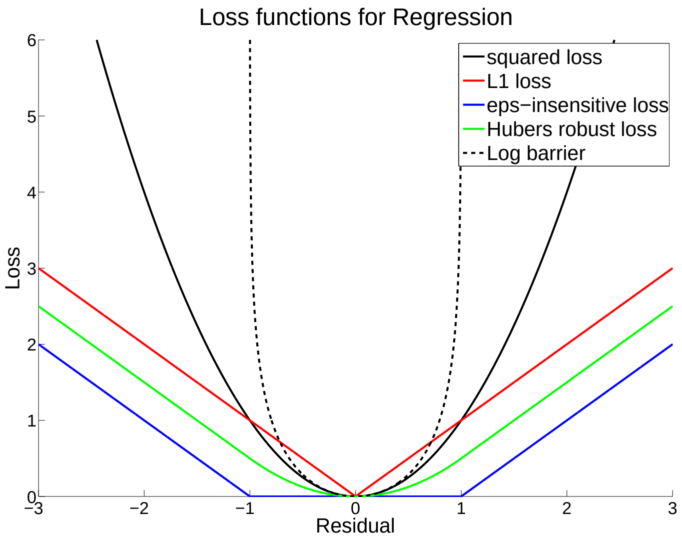
$$R(f) = \mathbb{E}\big[L(Y, f(X))\big] = \mathbb{E}_X\big[\mathbb{E}_{Y|X}[L(Y, f(X) \,|\, X]\big].$$

Usually: Loss function takes as argument $|y - f(x)|$.
$L(y, f(x)) = L(|y - f(x)|)$.

$\Longrightarrow$ there is no generic loss function as in classification

Squared loss: $L(y, f(x)) = (y - f(x))^2$	$f_L^*(x) = \mathbb{E}_Y[Y \vert X = x]$,
L_1 **- loss:** $L(y, f(x)) = \vert y - f(x) \vert$	$f_L^*(x) = \mathrm{Median}(Y \vert X = x)$,
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 (puzzle)

Loss functions for regression III



Median is more stable than the mean