Machine Learning

Prof. Barbara Caputo
Dip. Ingegneria Informatica, Automatica e Gestionale, Roma

SAPIENZA
UNIVERSITÀ DI ROMA

Useful info (1)

Teacher: Barbara Caputo,
www.dis.uniroma1.it/~barbaracaputo

Assistant: Fabio M. Carlucci,

Where/how to find us:
Email: caputo@dis.uniroma1.it,
fabiom.carlucci@dis.uniroma1.it

Office: B109 (BC), A003 (IK)

Q/A time: Fri 9:30-10:30/after the lectures (BC)
You can come at other times at your own risk!

Useful info (2)

Exam modality:

Homeworks (lab experiences) +oral

For AIML students, the modality will remain as last year (homeworks +written exam)

Course Web site:
https://sites.google.com/site/machinelearning20162017/

To get slides (after lecture) send and email to F. M. Carlucci for access to course Dropbox

Outline

Introduction and basics (a crash course on probability)
Bayes decision theory
Principal Component Analysis
Regression
Non-parametric methods: K-NN
Discriminative Methods: Perceptron, Neural Networks,
Deep learning,
SVM, Kernels
Learning Theory: regularization, risk minimization,
VC dimension
Unsupervised Learning: clustering,
semi-supervised learning
Probabilistic Representation and modeling:
graphical models, Bayes nets, HMM, Reinforcement learning,
topic models.

Outline for AIML students

Introduction and basics (a crash course on probability)
Bayes decision theory
Principal Component Analysis
Non-parametric methods: K-NN
Discriminative Methods: Perceptron, SVM, Kernels
Learning Theory: regularization, risk minimization,
Unsupervised Learning: clustering,

Parameter estimation: MLE, MAP

Estimating Probabilities

Flipping a Coin

I have a coin, if I flip it, what's the probability it will fall with the head up?

Let us flip it a few times to estimate the probability:

The estimated probability is: **3/5** "Frequency of heads"

Why???... and How good is this estimation???

45

MLE for Bernoulli distribution

Data, D =

$$D = \{X_i\}_{i=1}^{n}, \; X_i \in \{\mathrm{H}, \mathrm{T}\}$$

MLE for Bernoulli distribution

Data, D =

$$D = \{X_i\}_{i=1}^{n},\ X_i \in \{\mathrm{H}, \mathrm{T}\}$$

$P(\text{Heads}) = \theta,\ P(\text{Tails}) = 1\text{-}\theta$

Flips are **i.i.d.**:

- **Independent** events
  - **Identically distributed** according to Bernoulli distribution

MLE: Choose θ that maximizes the probability of observed data

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta)$$ Identically distributed

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1-\theta)$$ Identically distributed

$$= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

47

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1}(1-\theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H}(1-\theta)^{\alpha_T - 1}\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{J(\theta)}$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1}(1-\theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H}(1-\theta)^{\alpha_T - 1}\Big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\alpha_H(1-\theta) - \alpha_T \theta\big|_{\theta=\hat{\theta}_{\mathrm{MLE}}} = 0$$

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

48

What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

The Bayesian way…

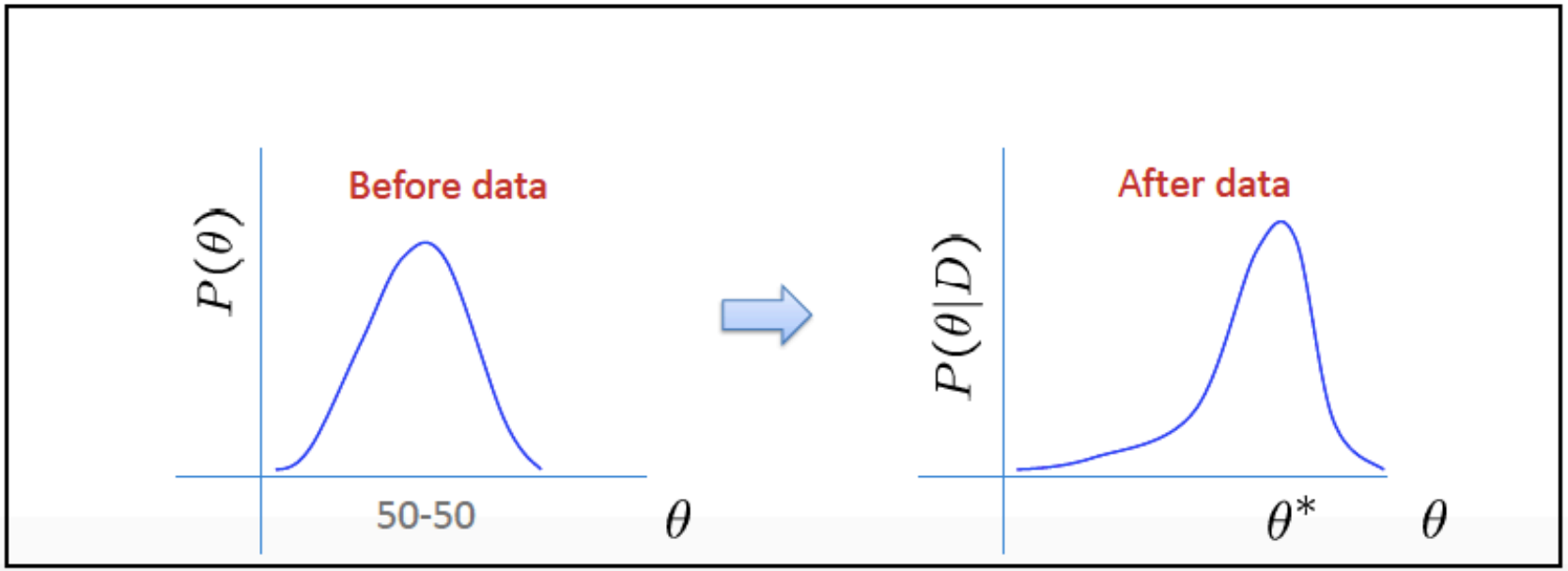
Rather than estimating a single θ, we obtain a distribution over possible values of θ

What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

The Bayesian way…

Rather than estimating a single θ, we obtain a distribution over possible values of θ



49

Bayesian Learning

- Use Bayes rule:

$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})}$$

- Or equivalently:

$$P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$$

posterior likelihood prior

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial $$P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial $$P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

If the prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial $P(\mathcal{D} \mid \theta) = \binom{n}{\alpha_H} \theta^{\alpha_H} (1-\theta)^{\alpha_T}$

If the prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

$\Rightarrow$ posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

P(θ) and P(θ| D) have the same form! [Conjugate prior]

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid D) = \arg\max_{\theta} P(D \mid \theta) P(\theta) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

51

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

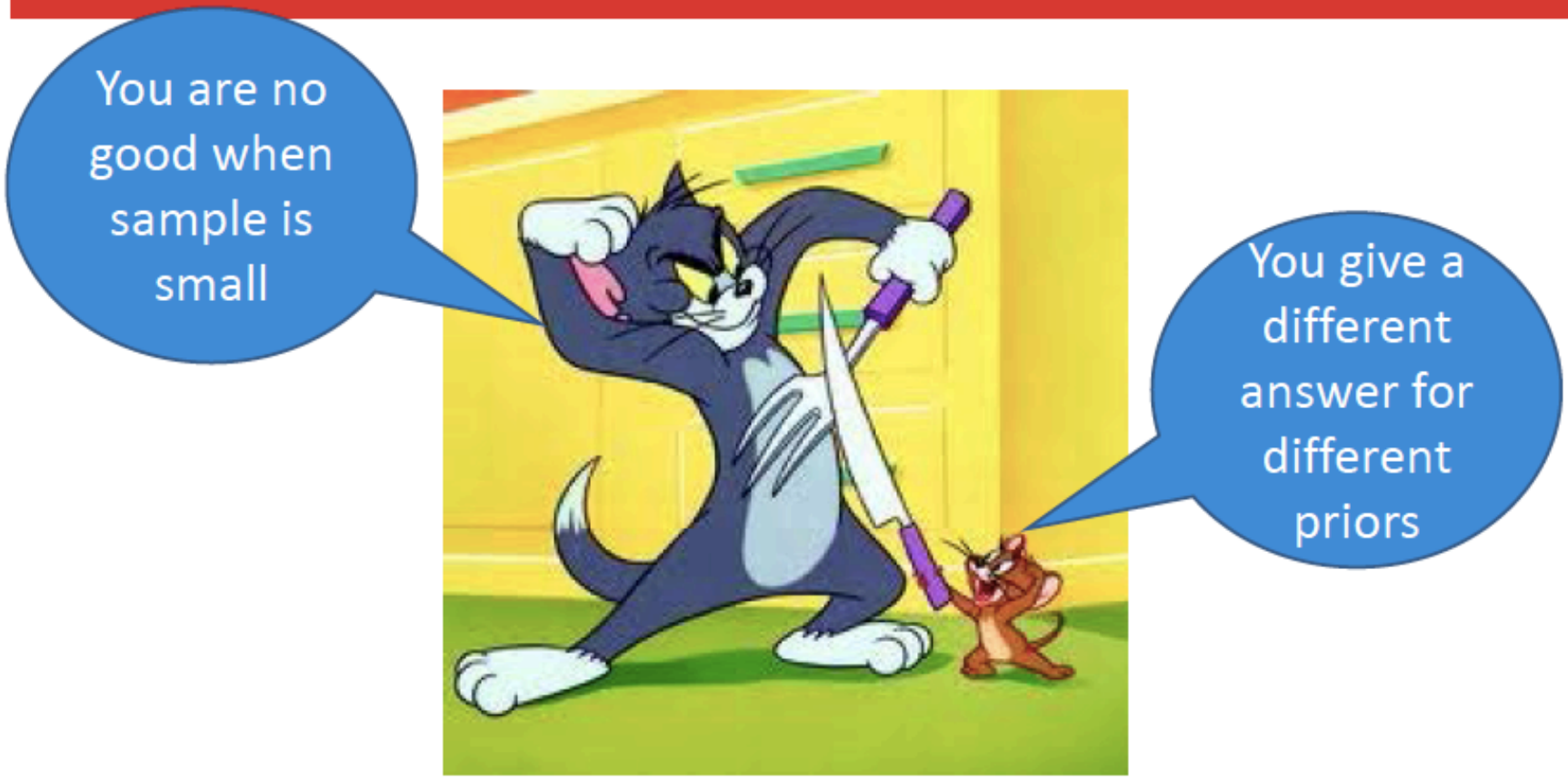
- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

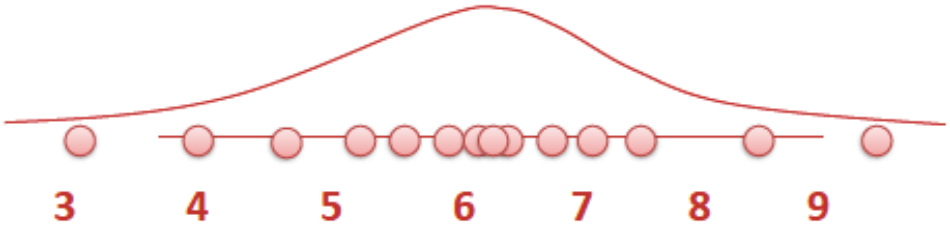
$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D)$$
$$= \arg\max_{\theta} P(D|\theta)P(\theta)$$

When is MAP same as MLE?

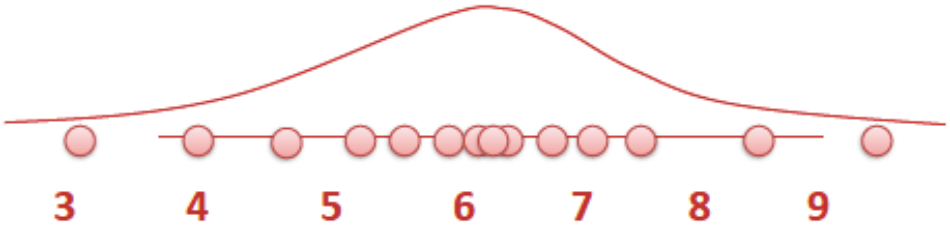
Bayesians vs.Frequentists



What about continuous features?

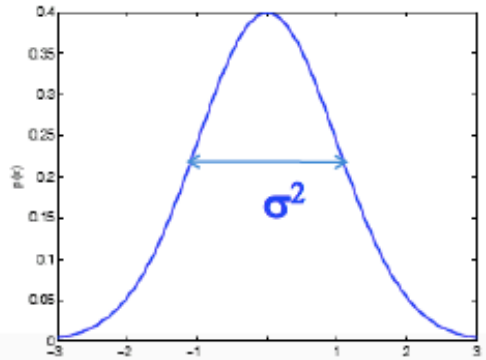


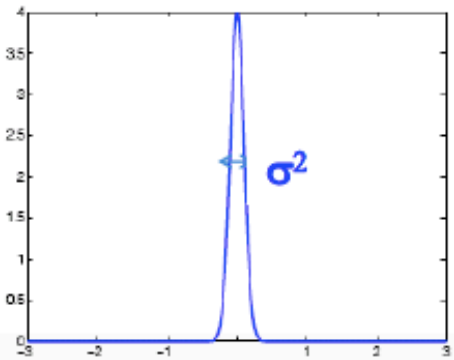
What about continuous features?



Let us try Gaussians…

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \mathcal{N}_x(\mu, \sigma)$$





54

MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

Independent draws

MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-(X_i-\mu)^2/2\sigma^2}$$ Identically distributed

MLE for Gaussian mean and variance

Choose $\theta = (\mu, \sigma^2)$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$ Independent draws

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-(X_i-\mu)^2/2\sigma^2}$$ Identically distributed

$$= \arg \max_{\theta=(\mu,\sigma^2)} \underbrace{\frac{1}{2\sigma^2} e^{-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2}}_{J(\theta)}$$

55

MLE for Gaussian mean and variance

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

MLE for Gaussian mean and variance

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Note: MLE for the variance of a Gaussian is **biased**

[Expected result of estimation is **not** the true parameter!]

Unbiased variance estimator: $\hat{\sigma}^2_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$

56

Five minutes break!

Bayes Decision Rule

x is an observation for which:

if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⟹ True state of nature = ω_1

if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⟹ True state of nature = ω_2

Bayes Decision Rule

x is an observation for which:

if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ⟹ True state of nature = ω_1
if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ⟹ True state of nature = ω_2

Therefore:

whenever we observe a particular x, the probability of error is :

$P(error \mid x) = P(\omega_1 \mid x)$ *if we decide* ω_2
$P(error \mid x) = P(\omega_2 \mid x)$ *if we decide* ω_1

Bayes Decision Rule minimizes probability of error

Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$;
otherwise decide ω_2

Bayes Decision Rule minimizes probability of error

Decide ω_1 *if* $P(\omega_1 \mid x) > P(\omega_2 \mid x)$;
otherwise decide ω_2

Therefore:

$$P(error \mid x) = min\ [\ P(\omega_1 \mid x),\ \ P(\omega_2 \mid x)\ ]$$

(Bayes decision)

Bayes Decision Theory – Continuous Features

- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing actions and not only decide on the state of nature
  - Introduce a loss of function which is more general than the probability of error

Loss Function

- Allowing actions other than classification primarily allows the possibility of rejection
- Refusing to make a decision in close or bad cases!
- The loss function states how costly each action taken is

Loss Function Definition

Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or "categories")

Loss Function Definition

Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or "categories")

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible actions

Loss Function Definition

Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or "categories")

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible actions

Let $\lambda(\alpha_i \mid \omega_j)$ be the loss incurred for taking

action α_i when the state of nature is ω_j

Overall Risk

R = Sum of all $R(\alpha_i \mid x)$ for $i = 1,\ldots,a$

Overall Risk

R = Sum of all $R(\alpha_i \mid x)$ for $i = 1,\ldots,a$

Conditional risk

Minimizing R ⟺ Minimizing $R(\alpha_i \mid x)$ for $i = 1,\ldots, a$

Expected Loss with action i

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$$

Overall Risk

R = Sum of all $R(\alpha_i \mid x)$ for i = 1,…,a

Conditional risk

Minimizing R ⟷ Minimizing $R(\alpha_i \mid x)$ for $i = 1, \ldots, a$

Expected Loss with action i

$$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

Select the action α_i for which $R(\alpha_i \mid x)$ is minimum

R is minimum and R in this case is called the Risk

Bayes risk = best performance that can be achieved

Two-category classification

α_1 : deciding ω_1

α_2 : deciding ω_2

$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$

Two-category classification

α_1 : deciding ω_1

α_2 : deciding ω_2

$\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$

loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

$$R(\alpha_1 \mid x) = \lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x)$$

Minimum Risk Decision Rule

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

action α_1: "decide ω_1" is taken

Minimum Risk Decision Rule

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

action α_1: "decide ω_1" is taken

This results in the equivalent rule :

decide ω_1 if:

$$(\lambda_{21} - \lambda_{11})\, P(x \mid \omega_1)\, P(\omega_1) > (\lambda_{12} - \lambda_{22})\, P(x \mid \omega_2)\, P(\omega_2)$$

and decide ω_2 otherwise

Minimum-Error-Rate Classification

- Actions are decisions on classes

  If action α_i is taken and the true state of nature is ω_j then: decision is correct if $i = j$ and in error if $i \neq j$

  Seek a decision rule that minimizes the *probability of error* which is the *error rate*

That’s all!