Intelligent Systems
Statistical Machine Learning

Carsten Rother, Dmitrij Schlesinger

WS2016/2017, 21.12.2016

Our model and tasks

The **model**: two variables are usually present:

- the first one is typically discrete $k \in K$ and is called “class”

- the second one is often continuous $x \in X$ and is called “observation”

The **recognition** (inference) task:
Let the joint probability distribution $p(x, k)$ be “given”.
Observe x, estimate k.

The (statistical) **learning** task: given a training set

$$L = \big((x_1, k_1), (x_2, k_2), \ldots, (x_l, k_l)\big)$$

“find” the corresponding probability distribution $p(x, k)$

Outline

- Decision making:
  - Bayesian Decision Theory
  - Non-Bayesian formulation

- Statistical Learning — Maximum Likelihood Principle

Bayesian Decision Theory: Idea — a game

Somebody samples a pair (x, k) according to a p.d. $p(x, k)$

He keeps k hidden and presents x to **you**

You decide for some k^* according to a chosen **decision strategy**

Somebody penalizes your decision according to a **loss-function**, i.e. he compares your decision with the "true" hidden k

---

You know both $p(x, k)$ and the loss-function (how does he compare)

Your goal is to design the decision strategy in order to pay as less as possible in average.

Bayesian Risk

The decision set D. Note: it needs not to coincide with K !!!
Examples: decisions “I don’t know”, “surely not this class” etc.

Decision strategy (mapping) $e : X \to D$

Loss-function $C : D \times K \to \mathbb{R}$

The **Bayesian Risk** is the expected loss:

$$R(e) = \sum_x \sum_k p(x,k) \cdot C\big(e(x),k\big) \to \min_e$$

(should be minimized with respect to the decision strategy).

For a particular observation x :

$$R(d) = \sum_k p(k|x) \cdot C(d,k) \to \min_d$$

Maximum A-posteriori Decision (MAP)

The loss is the simplest one (called delta-function):

$$C(k,k') = \delta(k \neq k') = \begin{cases} 1 & \text{if } \; k \neq k' \\ 0 & \text{otherwise} \end{cases}$$

i.e. we pay 1 if we are wrong, no matter what error do we make.

From that follows:

$$R(k) = \sum_{k'} p(k'|x) \cdot \delta(k \neq k') =$$

$$= \sum_{k'} p(k'|x) - p(k|x) = 1 - p(k|x) \rightarrow \min_k$$

$$p(k|x) \rightarrow \max_k$$

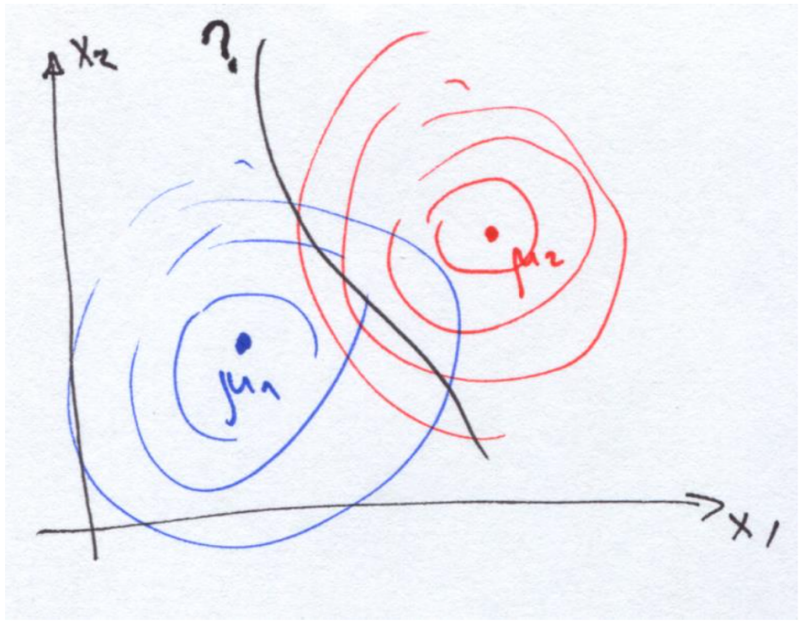
A MAP example

Let $K = \{1, 2\},\ x \in \mathbb{R}^2,\ p(k)$ be given. The conditional probability distribution for observations given classes are Gaussians:

$$p(x|k) = \frac{1}{2\pi\sigma_k^2} \exp\left[-\frac{||x - \mu_k||^2}{2\sigma_k^2}\right]$$

The loss function is $C(k, k') = \delta(k \neq k')$, i.e. we want MAP

The decision strategy (the mapping $e : X \to K$) **partitions** the input space into two regions: the one corresponding to the first and the one corresponding to the second class. How does this partition look like ?



A MAP example

For a particular x we decide for 1, if

$$p(1) \cdot \frac{1}{2\pi\sigma_1^2} \exp\left[-\frac{||x-\mu_1||^2}{2\sigma_1^2}\right] > p(2) \cdot \frac{1}{2\pi\sigma_2^2} \exp\left[-\frac{||x-\mu_2||^2}{2\sigma_2^2}\right]$$

Special case for simplicity: $\sigma_1 = \sigma_2$ and $p(1) = p(2)$

$\rightarrow$ the decision strategy is:

$$||x-\mu_1||^2 < ||x-\mu_2||^2$$

$$||x||^2 - 2\langle x, \mu_1\rangle + ||\mu_1||^2 < ||x||^2 - 2\langle x, \mu_2\rangle + ||\mu_2||^2$$

$$\langle x, \mu_2 - \mu_1\rangle < const$$

$\rightarrow$ a linear classifier — the hyperplane orthogonal to $\mu_2 - \mu_1$

(more examples at the exercise)

Decision with rejection

The decision set is $D = K \cup \{rw\}$, i.e. it is extended by a special decision “I don’t know” (rejection). The loss-function is

$$C(d,k) = \begin{cases} \delta(k \neq k') & \text{if } \; d \in K \\ \varepsilon & \text{if } \; d = rw \end{cases}$$

i.e. we pay a (reasonable) penalty if we are lazy to decide.

Case-by-case analysis:

1. We decide for a class $d \in K$, decision is MAP $d = k^*$, the loss for this is $1 - p(k^*|x)$

2. We decide to reject $d = rw$, the loss for this is ε

$\rightarrow$ The decision strategy: Compare $p(k^*|x)$ with $1 - \varepsilon$ and decide for the variant with the greater value.

Outline

- Decision making:
  - Bayesian Decision Theory
  - Non-Bayesian formulation

- Statistical Learning — Maximum Likelihood Principle

Non-Bayesian Decisions

Despite the generality of Bayesian approach, there are many tasks which cannot be expressed within the Bayesian framework:

- It is difficult to establish a penalty function, e.g. it does not assume values from the totally ordered set.

- A priori probabilities $p(k)$ are not known or cannot be known because k is not a random event.

An example – Russian fairy tales hero

When he turns to the left, he loses his horse, when he turns to the right, he loses his sword, and if he turns back, he loses his beloved girl.

Is the sum of p_1 horses and p_2 swords is less or more than p_3 beloved girls ?

Example: decision while curing a patient

We have:

$x \in X$ — observations (features) measured on a patient

$k \in K = \{healthy, seriously\ sick\}$ — hidden states

$d \in D = \{do\ not\ cure, apply\ a\ drug\}$ — decisions

Penalty problem:

how to assign real number to a penalty?

	do not cure	*apply a drug*
healthy	Correct decision	small health damage
seriously sick	death possible	Correct decision

Example: enemy or allied airplane ?

Observation x describes the observed airplane.

Two hidden states: $\begin{cases} k = 1 & allied\ airplane \\ k = 2 & enemy\ airplane \end{cases}$

The conditional probability $p(x|k)$ can depend on the observation x in a complicated manner but **it exists** and describes dependencies of the observation x on the situation k correctly.

A-priori probabilities $p(k)$ are not known and can not be known in principle.

$\rightarrow$ the hidden state k is **not a random event**

Neyman-Pearson Task (1928, 1933)

The strategy (partitioning the input space into two subsets $X_1 \cup X_2 = X$) is characterized by two numbers:

1. "Probability" of the **false positive** (false alarm)

$$\omega(1) = \sum_{x \in X_2} p(x|1)$$

2. "Probability" of the **false negative** (overlooked danger)

$$\omega(2) = \sum_{x \in X_1} p(x|2)$$

Minimize the conditional probability of the false positive subject to the condition that the false negative is bounded:

$$\omega(1) \rightarrow \min_{X_1, X_2} \quad \text{s.t.} \;\; \omega(2) \leq \varepsilon$$

Outline

- Decision making:
  - Bayesian Decision Theory
  - Non-Bayesian formulation

- Statistical Learning — Maximum Likelihood Principle

Learning

Let a **parameterized class** (family) of probability distributions be given, i.e. $p(x;\theta) \in \mathcal{P}$

Example — the set of Gaussians in $\mathbb{R}^n$

$$p(x;\mu,\sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{||x-\mu||^2}{2\sigma^2}\right]$$

parameterized by the mean $\mu \in \mathbb{R}^n$ and standard deviation $\sigma \in \mathbb{R}$

Let the **training data** be given: $L = (x_1, x_2, \ldots, x_{|L|})$

One have to decide for a particular probability distribution from the given family, i.e. for a particular parameter (e.g. $\theta^* = (\mu^*, \sigma^*)$) for Gaussians).

Maximum Likelihood Principle

Assumption: the training dataset is a realization of the unknown probability distribution – it is sampled according to it.

$\rightarrow$ What is observed should have a high probability

$\rightarrow$ Maximize the probability of the training data with respect to the unknown parameter

$$p(L;\theta) \rightarrow \max_{\theta}$$

Example: Gaussian

We have a Family of probability distributions for $x \in \mathbb{R}^n$:

$$p(x; \mu, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{||x-\mu||^2}{2\sigma^2}\right]$$

and a training set $L = (x_1, x_2, \ldots, x_{|L|})$

Assumption: the samples are **independent identically distributed** (iid.), i.e. the probability of the training set is

$$p(L; \mu, \sigma) = \prod_l p(x_l; \mu, \sigma)$$

Take logarithm and maximize it with respect to the unknown parameters:

$$\ln p(L; \mu, \sigma) = \sum_l \ln p(x_l; \mu, \sigma) \rightarrow \max_{\mu,\sigma}$$

Example: Gaussian

Substitute the model:

$$\sum_l \left[-n \ln \sigma - \frac{||x_l - \mu||^2}{2\sigma^2} \right] = -|L| \cdot n \cdot \ln \sigma - \frac{1}{2\sigma^2} \sum_l ||x_l - \mu||^2 \rightarrow \max_{\mu,\sigma}$$

Assume, we are interested only in the center μ, i.e. σ is given. Then the problem can be further simplified to

$$\sum_l ||x_l - \mu||^2 \rightarrow \min_{\mu}$$

Take the derivative, set it to zero

$$\frac{\partial}{\partial \mu} = \sum_l (x_l - \mu) = \sum_l x_l - |L| \cdot \mu = 0$$

and resolve

$$\mu = \frac{\sum_l x_l}{|L|}$$

i.e. the mean value over the dataset.

Some comments

Maximum Likelihood estimator is not the only estimator – there are many others as well.

Maximum Likelihood is **consistent**, i.e. it gives the true parameters for infinite training sets.

Consider the following experiment for an estimator:
1. We generate **infinite** numbers of training sets each one being **finite**;
2. For each training set we estimate the parameter;
3. We **average** all estimated values.

If the average is the true parameter, the estimator is called **unbiased**.

Maximum Likelihood is not always unbiased – it depends on the parameter to be estimated. Examples – mean for a Gaussian is unbiased, standard deviation – not.

Summary

Today: statistical Machine Learning

1. The model: two kinds of random variables — observations, hidden variables. Tasks: inference and learning

2. Decision making:

   1) Bayesian formulations — decision strategies, Bayesian risk, MAP and others

   2) Non-Bayesian formulations — Neyman-Pearson Task

3. Learning: Maximum Likelihood Principle, consistency, (un)biased estimators