Machine Learning
Bayesian Decision Theory

Dmitrij Schlesinger

WS2013/2014, November 5, 2013

COMPUTER VISION LAB DRESDEN

TECHNISCHE UNIVERSITÄT DRESDEN

Recognition

The model:

Let two random variables be given:

- 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"

Let the joint probability distribution $p(x, k)$ be "given"

As k is discrete it is often specified by $p(x, k) = p(k) \cdot p(x|k)$

The recognition task: given x, estimate k

Usual problems (questions):

- How to estimate k from x ? (today)
- The joint probability is not always explicitly specified
- The set K is sometimes huge

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 to 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

Notations:

The **decision set** D. Note: it needs not to coincide with K !!!
Examples: decisions like "I don't know", "not this class" ...

Decision strategy is a mapping $e : X \to D$

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

The **Bayesian Risk** of a strategy e is the expected loss:

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

It should be minimized with respect to the decision strategy

Some variants

General:

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

Almost always:
decisions can be made for different x independently (the set of decision strategies is not restricted). Then:

$$R\Big(e(x)\Big) = \sum_k p(x,k) \cdot C\Big(e(x), k\Big) \to \min_{e(x)}$$

Very often: the decision set coincides with the set of classes, i.e. $D = K$

$$\begin{aligned} k^* &= \arg\min_k \sum_{k'} p(x,k') \cdot C(k,k') = \\ &= \arg\min_k \sum_{k'} p(k'|x) \cdot C(k,k') \end{aligned}$$

Maximum A-posteriori Decision (MAP)

The Loss is the simplest one:

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

i.e. we pay 1 if the answer is not the true class, no matter what error we make.

From that follows:

$$\begin{aligned} R(k) &= \sum_{k'} p(k'|x) \cdot \delta(k \neq k') = \\ &= \sum_{k'} p(k'|x) - p(k|x) = 1 - p(k|x) \to \min_k \end{aligned}$$

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

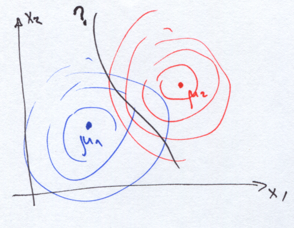
A MAP example

Let $K = \{1, 2\}$, $x \in \mathbb{R}^2$, $p(k)$ be given. Conditional probability distributions 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 $\delta(k \neq k')$, i.e. we want MAP.

The decision strategy $e : X \to K$ 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) $\sigma_1 = \sigma_2$
$\rightarrow$ the decision strategy is (derivation on the board)

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

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

More classes, equal σ and $p(k)$ $\rightarrow$ Voronoi-diagram
More classes, equal σ, different $p(k)$ $\rightarrow$ Fischer-classifier
Two classes, different σ – a general quadratic curve
etc.

Decision with rejection

The decision set is $D = K \cup \{r\}$, i.e. extended by a special decision "I don't know". The loss-function is

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

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

Case-by-case analysis:

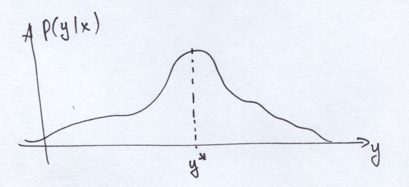
1. We decide for a class $d \in K$, then the decision is MAP $d = k^* = \arg\max_k p(k|x)$, the loss for this is $1 - p(k^*|x)$
2. We decide to reject $d = r$ and pay ε for this

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

Other simple loss-functions

Let the set of classes be **structured** (in some sense)

Example:
We have a probability density $p(x, y)$ with an observations x and a **continuous** hidden value y. Suppose, we know $p(y|x)$ for a given x, for which we would like to infer y.



The Bayesian Risk reads:

$$R\Big(e(x)\Big) = \int_{-\infty}^{\infty} p(y|x) \cdot C\Big(e(x), y\Big) dy$$

Other simple loss-functions

Simple δ-loss-function $\rightarrow$ MAP (not interesting anymore)

Loss may account for **differences** between the decision and the "true" hidden value, for instance $C(d, y) = (d - y)^2$, i.e. we pay depending on the **distance**.

Than (see board again):

$$\begin{aligned} e(x) &= \arg\min_{d} \int_{-\infty}^{\infty} p(y|x) \cdot (d - y)^2 dy = \\ &= \int_{-\infty}^{\infty} y \cdot p(y|x) dy = \mathbb{E}_{p(y|x)}[y] \end{aligned}$$

Other choices: $C(d, y) = |d - y|$, $C(d, y) = \delta(|d - y| > \varepsilon)$, combination with "rejection" etc.

Additive loss-functions – an example

	Q_1	Q_2	$\dots$	Q_n
P_1	1	0	$\dots$	1
P_2	0	1	$\dots$	0
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
P_m	0	1	$\dots$	0
"$\sum$"	?	?	$\dots$	?

Consider a "questionnaire": m persons answer n questions. Furthermore, let us assume that persons are rated – a "reliability" measure is assigned to each one.

The goal is to find the "right" answers for all questions.

Strategy 1:
Choose the **best** person and take **all** his/her answers.

Strategy 2:

- Consider a particular question
- Look, what **all** the people say concerning this, do (weighted) voting

Additive loss-functions – example interpretation

People are classes k, reliability measure is the posterior $p(k|x)$

Specialty:
classes consist of "parts" (questions) – classes are **structured**

The set of classes is $k = (k_1, k_2 \dots k_m) \in K^m$, it can be seen as a vector of m components each one being a simple answer (0 or 1 in the above example)

The "Strategy 1" is MAP

How to derive (consider, understand) the other decision strategy from the viewpoint of the Bayesian Decision Theory?

Additive loss-functions

Consider the simple $C(k,k')=\delta(k\neq k')$ loss for the case classes are structured – it does not reflect **how strong** the class and the decision disagree

A better (?) choice – additive loss-function

$$C(k,k')=\sum_i c_i(k_i,k'_i)$$

i.e. disagreements of all components are summed up

Substitute it in the formula for Bayesian Risk, derive and look what happens ...

Additive loss-functions – derivation

$$\begin{aligned} R(k) &= \sum_{k'} \Big[p(k'|x) \cdot \sum_i c_i(k_i, k'_i) \Big] = / \text{ swap summations} \\ &= \sum_i \sum_{k'} c_i(k_i, k'_i) \cdot p(k'|x) = / \text{ split summation} \\ &= \sum_i \sum_{l \in K} \sum_{k' : k'_i = l} c_i(k_i, l) \cdot p(k'|x) = / \text{ factor out} \\ &= \sum_i \sum_{l \in K} \Big[c_i(k_i, l) \cdot \sum_{k' : k'_i = l} p(k'|x) \Big] = / \text{ red are marginals} \\ &= \sum_i \sum_{l \in K} c_i(k_i, l) \cdot p(k'_i = l | x) \to \min_k \end{aligned}$$

/ independent problems

$$\Rightarrow \quad \sum_{l \in K} c_i(k_i, l) \cdot p(k'_i = l | x) \to \min_{k_i} \quad \forall i$$

Additive loss-functions – the strategy

1. Compute **marginal** probability distributions for values

$$p(k'_i{=}l|x) = \sum_{k':k'_i=l} p(k'|x)$$

   for each variable i and each value l

2. Decide for each variable "independently" according to its marginal p.d. and the local loss c_i

$$\sum_{l \in K} c_i(k_i, l) \cdot p(k'_i{=}l|x) \rightarrow \min_{k_i}$$

   This is again a Bayesian Decision Problem – minimize the average loss

Additive loss-functions – a special case

For each variable we pay 1 if we are wrong:

$$c_i(k_i, k'_i) = \delta(k_i \neq k'_i)$$

The overall loss is the number of misclassified variables (wrongly answered questions)

$$C(k, k') = \sum_i \delta(k_i \neq k'_i)$$

and is called **Hamming distance**

The decision strategy is **Maximum Marginal Decision**

$$k^*_i = \arg\max_l p(k'_i{=}l|x) \quad \forall i$$