Data analysis: Statistical principals and computational methods

Inference in MRF-s

Dmitrij Schlesinger, Carsten Rother

SS2014, 25.06.2014

COMPUTER VISION LAB DRESDEN

TECHNISCHE UNIVERSITÄT DRESDEN

Inference (Recognition)

The model:
Let two random variables be given:

- The first one is typically discrete ($k \in K$) – “class”
- The second one is arbitrary ($x \in X$) – “observation”

Let the joint probability distribution $p(x, k)$ be “known”

The recognition task: given x, estimate k
Usual problems (questions):

- How to estimate k from x ?
  $\rightarrow$ **Bayesian Decision Theory**
- The joint probability is not always explicitly specified
- The set K is sometimes huge,
  e.g. the set of all labelings in MRF

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

Another "writing style":

$$d^*(x) = \arg\min_d \sum_k p(k|x) \cdot C(d,k)$$

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:

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

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

i.e. choose the value with the highest a-posteriori probability

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 $k = (k_1, k_2 \ldots k_m) \in K^m$, 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 that 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

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

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

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

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**

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

Minimum Marginal Square Error (MMSE)

Assume, the values l for k_i are numbers (vectors)

Examples:

- in Tracking it is the set of all possible positions of the object to be tracked
- in Stereo it is the set of all disparity/depth values etc.

$\rightarrow$ a more reasonable (additive) loss should account for **metric** difference between the decision and the true position, e.g.

$$C(k,k') = \sum_i c_i(k_i,k'_i) = \sum_i \|k_i - k'_i\|^2$$

The task to be solved for each position i is

$$\sum_{l \in K} \|k_i - l\|^2 \cdot p(k'_i = l|x) \rightarrow \min_{k_i}$$

Minimum Marginal Square Error (MMSE)

$$\sum_{l \in K} \|k_i - l\|^2 \cdot p(k'_i{=}l|x) \to \min_{k_i}$$

$$\frac{\partial}{\partial k_i} = \sum_{l \in K} 2 \cdot (k_i - l) \cdot p(k'_i{=}l|x) = 0$$

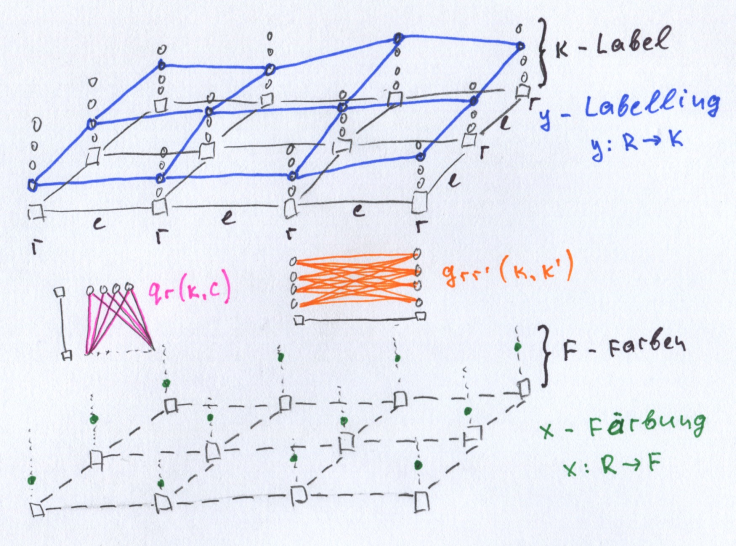
$$\sum_{l \in K} k_i \cdot p(k'_i{=}l|x) = \sum_{l \in K} l \cdot p(k'_i{=}l|x)$$

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

The optimal decision for i-th variable is the expectation (average) in the corresponding marginal probability distribution

Note: the decision is not necessarily an element of K, e.g. it may be real-valued $\to$ sets D and K are different.

Back to MRF-s



Back to MRF-s

Graph $G = (V, \mathcal{E})$, K – label set, F – observation set
$y \in \mathcal{Y} : V \to K$ – labeling, $x \in \mathcal{X} : V \to F$ – observation

An elamantary event is a pair (x, y). Its (negative) energy:

$$E(x,y) = \sum_{ij \in \mathcal{E}} \psi_{ij}(y_i, y_j) + \sum_{i \in V} \psi_i(x_i, y_i)$$

Its probability:

$$p(x,y) = \frac{1}{Z} \exp\Big[-E(x,y)\Big]$$

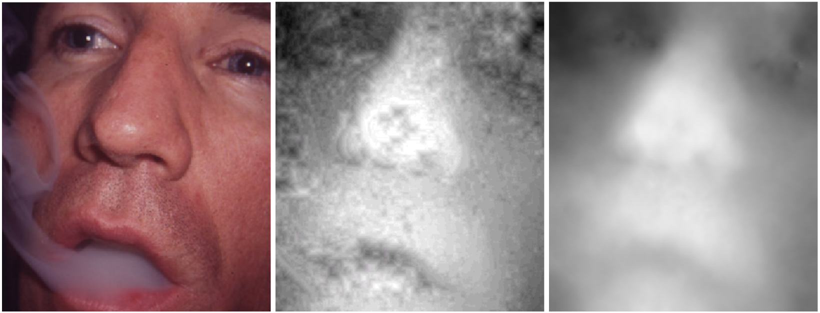
With the partition function:

$$Z = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \exp\Big[-E(x,y)\Big]$$

Note: MAP for MRF-s is Energy Minimization !!!

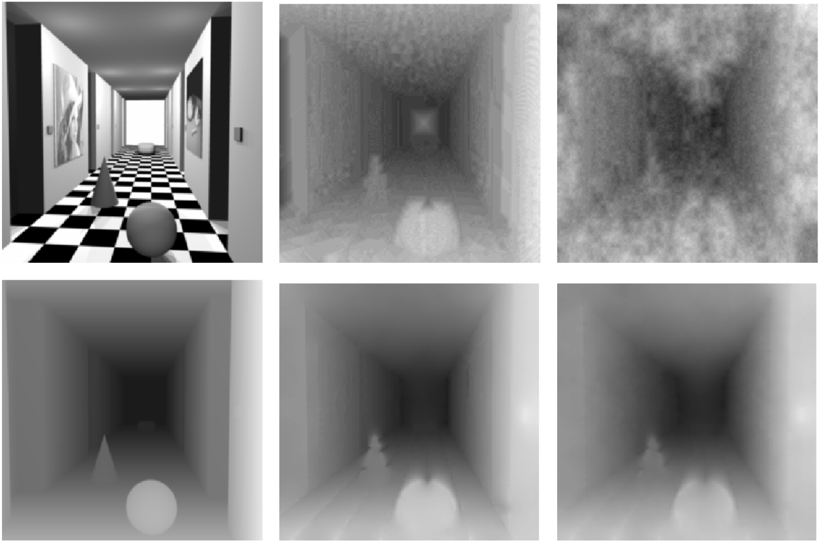
Example for MMSE – Stereo

MAP vs. MMSE



The left image MAP MMSE

Example for MMSE – Stereo



Other examples

Denoising:
Uwe Schmidt, Qi Gao, and Stefan Roth. *A generative perspective on MRFs in low-level vision*. CVPR 2010

Deconvolution:
Uwe Schmidt, Kevin Schelten, and Stefan Roth. *Bayesian deblurring with integrated noise estimation*. CVPR 2011

Segmentation:
remember on demo

How to estimate marginal label probability distributions (NP in general)? $\rightarrow$ sampling (later, will also be needed for learning)

Summary

Before:

- Markov chains
- Energy minimization

Today:

- Bayesian Decision Theory
- Additive loss-functions – structural loss
- MMSE for MRF-s

Next classes:

- Statistical learning (Maximum Likelihood)
- Discriminative learning (Structural SVM)