Data analysis: Statistical principals and computational methods

Statistical Learning in MRF-s

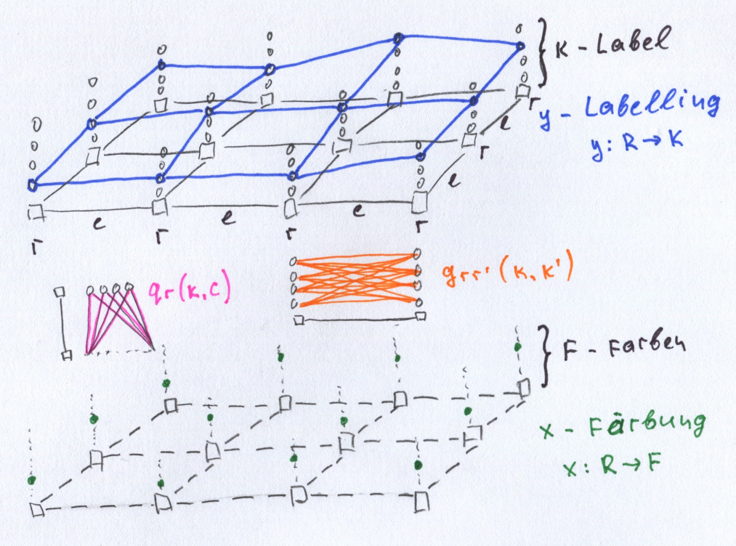
Dmitrij Schlesinger, Carsten Rother

SS2014, 02.07.2014

COMPUTER VISION LAB DRESDEN

TECHNISCHE UNIVERSITÄT DRESDEN

Remember the model



Remember the model

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 elemantary 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:

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

Remember the inference with an additive loss

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

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

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

Remember the "question"

How to compute the marginal probability distributions

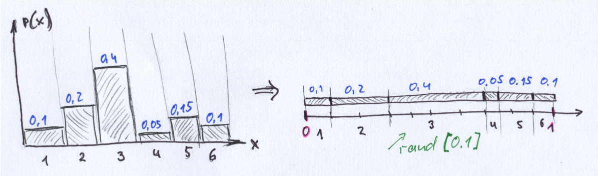
$$p(y_i{=}l|x) = \sum_{y:y_i=l} p(k|x)$$

It is not necessary to eat up the whole kettle completely in order to test a soup. It is often enough to stir it carefully and take just a spoon.

The idea: instead to sum over **all** labelings, **sample** a couple of them according to the target probability distribution and average $\rightarrow$ the **probabilities** are substituted by the relative **frequencies**

Sampling

Example: the values of a discrete Variable $x \in \{1,2,3,4,5,6\}$ have to be drawn from $p(x) = (0.1, 0.2, 0.4, 0.05, 0.15, 0.1)$



The algorithm: input – $p(x)$, output – a sample from $p(x)$

$a[1] = p[1]$

for $i{=}2$ `bis` n

$\quad a[i] = a[i{-}1] + p[i]$

$r = rand[0,1]$

for $i = 1$ `bis` n

$\quad$ **if** $a[i] > r$ **return** i

Gibbs Sampling

Task – draw an $x = (x_1, x_2 \dots x_m)$ (vector) from $p(x)$
Problem: $p(x)$ is not given explicitly

The way out:

- start with an arbitrary x^0
- sample the new one x^{t+1} "component-wise" from **conditional** probability distributions $p(x_i | x_1^t \dots x_{i-1}^t, x_{i+1}^t \dots x_m^t)$
- repeat it for all components i (Komponenten) many times

After such a sampling procedure (under some mild conditions):

- x^n does not depend on x^0
- x^n follows the target probability distribution $p(x)$

Gibbs Sampling

In MRF-s the conditional probability distributions can be easily computed !!!

The Markovian property

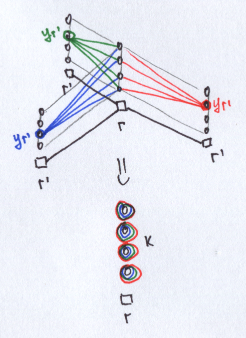
$$p(y_i|y_{V\setminus i}) = p(y_i|y_{N(i)})$$

(i.e. under the condition that the labels in the neighbouring nodes are fixed, $N(i)$ – neighbourhood structure) leads to

$$p(y_i{=}k|y_{N(i)}) \propto \exp\Big[-\psi_i(k) - \sum_{j\in N(i)} \psi_{ij}(k, y_j)\Big]$$

Gibbs Sampling

A relation to Iterated Conditional Modes:



– ICM considers the "conditional energies"

$$E_i(k) = \psi_i(k) + \sum_{j \in N(i)} \psi_{ij}(k, y_j)$$

and decides for the **best** label

– Gibbs Sampling **draws** new labels according to the conditional probabilities

$$p(y_i{=}k | y_{N(i)}) \propto \exp\Big[-E_i(k)\Big]$$

Maximum Likelihood for MRF-s (supervised)

The Model – no hidden variables, the energy is parameterized by a parameter θ to be learned:

$$p(y) = \frac{1}{Z(\theta)} \exp\Big[-E(y;\theta)\Big] \quad \text{with } Z(\theta) = \sum_y \exp\Big[-E(y;\theta)\Big]$$

Let a training set $L = (y^1, y^2 \dots y^{|L|})$ be given.

The Maximum Likelihood reads:

$$p(L;\theta) = \prod_l p(y^l;\theta) = \prod_l \frac{1}{Z(\theta)} \exp\Big[-E(y^l;\theta)\Big] \rightarrow \max_\theta$$

Take the logarithm:

$$F(\theta) = \ln p(L;\theta) = \sum_l \Big[-E(y^l;\theta) - \ln Z(\theta)\Big] =$$

$$= -\sum_l E(y^l;\theta) - |L| \cdot \ln Z(\theta) \rightarrow \max_\theta$$

Maximum Likelihood for MRF-s (supervised)

Consider the derivative with respect to θ (the gradient)

$$\frac{\partial F(\theta)}{\partial \theta} = -\sum_l \frac{\partial E(y^l;\theta)}{\partial \theta} - |L| \cdot \frac{\partial \ln Z(\theta)}{\partial \theta}$$

Apply the chain rule for the second addent:

$$\frac{\partial \ln Z(\theta)}{\partial \theta} = \frac{1}{Z(\theta)} \sum_y \exp\Big[-E(y;\theta)\Big] \cdot -\frac{\partial E(y;\theta)}{\partial \theta} =$$

$$= -\sum_y \frac{1}{Z(\theta)} \exp\Big[-E(y;\theta)\Big] \cdot \frac{\partial E(y;\theta)}{\partial \theta} =$$

$$= -\sum_y p(y;\theta) \cdot \frac{\partial E(y;\theta)}{\partial \theta}$$

Maximum Likelihood for MRF-s (supervised)

All together (the complete normalized gradient)

$$\frac{\partial F(\theta)}{\partial \theta} = -\frac{1}{|L|}\sum_{l} \frac{\partial E(y^l;\theta)}{\partial \theta} + \sum_{y} p(y;\theta) \cdot \frac{\partial E(y;\theta)}{\partial \theta}$$

The gradient is the difference of two **expectations**:

$$\frac{\partial F(\theta)}{\partial \theta} = -\mathbb{E}_{data}\left[\frac{\partial E(y;\theta)}{\partial \theta}\right] + \mathbb{E}_{model}\left[\frac{\partial E(y;\theta)}{\partial \theta}\right]$$

one over the training set and other over all elementary events.

The first one is called **data statistics** the second one is the **model statistics**.

Maximum Likelihood for MRF-s (supervised)

What is $\partial E(y;\theta)/\partial\theta$?

Example: let the unknown parameter θ is composed of unknown pairwise potentials $\psi_{ij}(k,k')$ (tables for all edges)
Consider a particular edge (i,j) and a label pair (k,k')

$$\frac{\partial E(y;\psi)}{\partial \psi_{ij}(k,k')} = \begin{cases} 1 & \text{if } \; y_i = k, y_j = k' \\ 0 & \text{otherwise} \end{cases}$$

It follows:

$$\frac{1}{|L|}\sum_l \frac{\partial E(y;\psi)}{\partial \psi_{ij}(k,k')} = n_{ij}(k,k')$$

$$\sum_y p(y;\psi)\cdot\frac{\partial E(y;\psi)}{\partial \psi_{ij}(k,k')} = p(y_i{=}k, y_j{=}k';\psi)$$

the first addend is the frequencies in the training set
the second one is the corresponding marginal probability

Maximum Likelihood for MRF-s (supervised)

To summarize (for the example, where ψ are learned)

Algorithm:

1. Compute $n_{ij}(k, k')$ from the training set
2. Repeat until convergence:
   a) Estimate the current marginal probabilities $p(y_i{=}k, y_j{=}k'; \psi)$ (e.g. by Gibbs Sampling)
   b) Compute the gradient as $p(y_i{=}k, y_j{=}k'; \psi) - n_{ij}(k, k')$ and apply it with a small step size

---

Further topics: supervised learning for hidden MRF-s, unsupervised learning (by gradient ascent, Expectation Maximization), conditional likelihood (the next lecture) etc.