Machine Learning
Discriminative Learning

Dmitrij Schlesinger

WS2013/2014, 15.11.2013

Discriminative models

There exists a joint probability distribution $p(x, k; \theta)$ (observation, class; parameter). The task is to learn θ

On the other side (see the "Bayesian Decision theory"),

$$R(d) = \sum_k p(k|x; \theta) \cdot C(d, k)$$

i.e. only the posterior $p(k|x; \theta)$ is relevant for the recognition.

The Idea: decompose the joint probability distribution into

$$p(x, k; \theta) = p(x) \cdot p(k|x; \theta)$$

with an **arbitrary** $p(x)$ and a **parameterized posterior**.

$\rightarrow$ learn the parameters of the posterior p.d. directly

Maximum Likelihood

Let the (complete) training data $L = \Big((x^l, k^l) \ldots\Big)$ be given.

$$p(L;\theta) = \prod_l \Big[p(x^l) \cdot p(k^l|x^l;\theta)\Big]$$

$$\ln p(L;\theta) = \sum_l \ln p(x^l) + \sum_l \ln p(k^l|x^l;\theta)$$

The first term can be omitted as we are not interested in $p(x)$

The second term is often called the **conditional likelihood**.

Maximum Likelihood Example

1. We consider a joint probability distribution $p(x, k) = p(k) \cdot p(x|k)$
2. We derive the posterior $p(k|x)$, i.e. we represent the joint p.d. as $p(x, k) = p(x) \cdot p(k|x)$
3. We forget $p(x)$ (assume that it is arbitrary) – we enlarge the family of considered p.d.-s
4. We look, how the Maximum Likelihood looks like

Example: two Gaussians of equal variance, i.e. $k \in \{1, 2\}$, $x \in \mathbb{R}^n$,

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

Maximum Likelihood Example

Posterior:

$$p(k{=}1|x) = \frac{p(1)p(x|1)}{p(1)p(x|1) + p(2)p(x|2)} = \frac{1}{1 + \frac{p(2)p(x|2)}{p(1)p(x|1)}} =$$

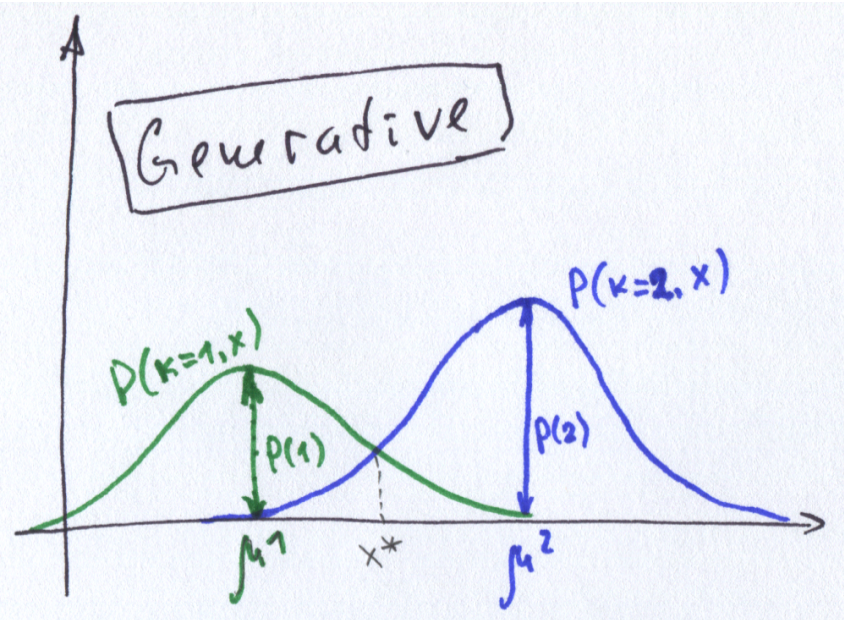
$$= \frac{1}{1 + \exp\left[-\frac{\|x-\mu^2\|^2}{2\sigma^2} + \frac{\|x-\mu^1\|^2}{2\sigma^2} + \ln p(2) - \ln p(1)\right]} =$$

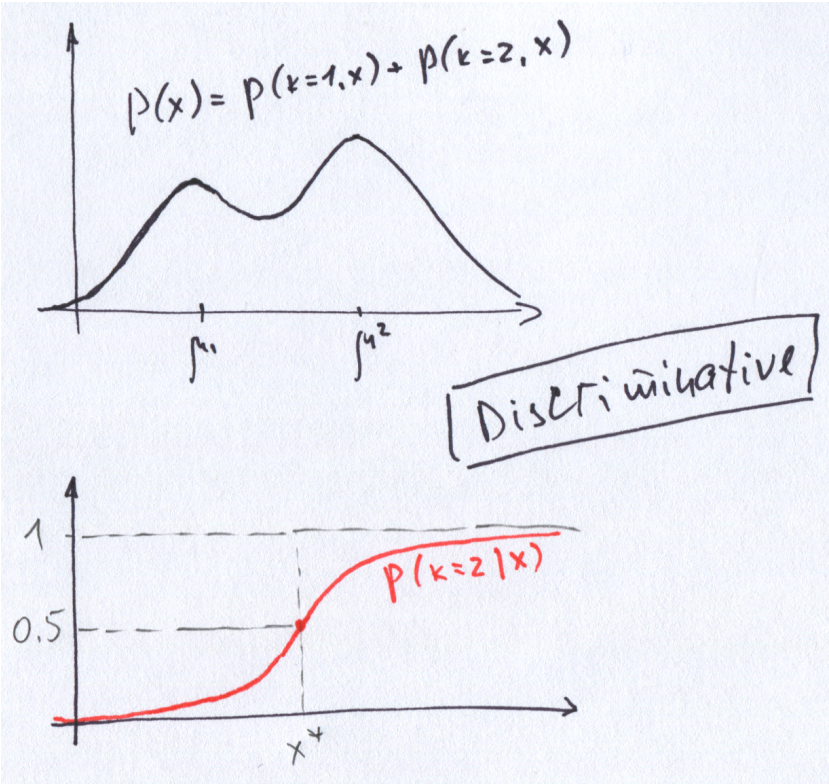
$$= \frac{1}{1 + \exp\Big(\langle x, w\rangle + b\Big)} \quad \text{with} \quad w = (\mu^2 - \mu^1)/\sigma^2$$

$$p(k{=}2|x) = 1 - p(k{=}1|x) = \frac{\exp\Big(\langle x, w\rangle + b\Big)}{1 + \exp\Big(\langle x, w\rangle + b\Big)}$$

Logistic regression model

Maximum Likelihood Example





Maximum Likelihood Example

Logistic regression (scalar products as simple multiplications):

$$p(k{=}1|x) = \frac{1}{1+\exp(wx+b)}, \quad p(k{=}2|x) = \frac{\exp(wx+b)}{1+\exp(wx+b)}$$

Conditional likelihood:

$$\begin{aligned} CL &= \sum_l \ln p(k^l|x^l;w,b) = \\ &= \sum_{l:k^l=1} -\ln\Big(1+\exp(wx^l+b)\Big) + \\ &+ \sum_{l:k^l=2} \Big[wx^l+b-\ln\Big(1+\exp(wx^l+b)\Big)\Big] = \\ &= w\cdot\sum_{l:k^l=2} x^l + b\cdot n_2 - \sum_l \ln\Big(1+\exp(wx^l+b)\Big) \to \max_{w,b} \end{aligned}$$

Maximum Likelihood Example

Gradient:

$$\frac{\partial CL}{\partial w} = \sum_{l:k^l=2} x^l - \sum_l \frac{\exp(wx^l + b)}{1 + \exp(wx^l + b)} x^l =$$

$$= \sum_{l:k^l=2} x^l - \sum_l p(k{=}2|x^l; w, b) x^l$$

$$\frac{\partial CL}{\partial b} = n_2 - \sum_l p(k{=}2|x^l; w, b)$$

It is not possible to resolve it analytically :-(

Note: the subject is concave $\rightarrow$ Gradient-method leads to the global solution :-)

Generative vs. discriminative

Posterior p.d.-s have less free parameters as joint ones

Compare (for Gaussians):

- $2n + 2$ free parameters for the generative representation $p(k, x) = p(k) \cdot p(x|k)$, i.e. $p(1)$, σ, μ^1, μ^2
- $n + 1$ free parameters for the posterior $p(k|x)$, i.e. w and b

$\rightarrow$ one posterior corresponds to many joint p.d.-s

Gaussian example again:
centers μ^1 and μ^2 are not relevant, but their difference $\mu^2 - \mu^1$ (see the board for the explanation).

Generative vs. discriminative

Consider two learning schemes for Gaussians:

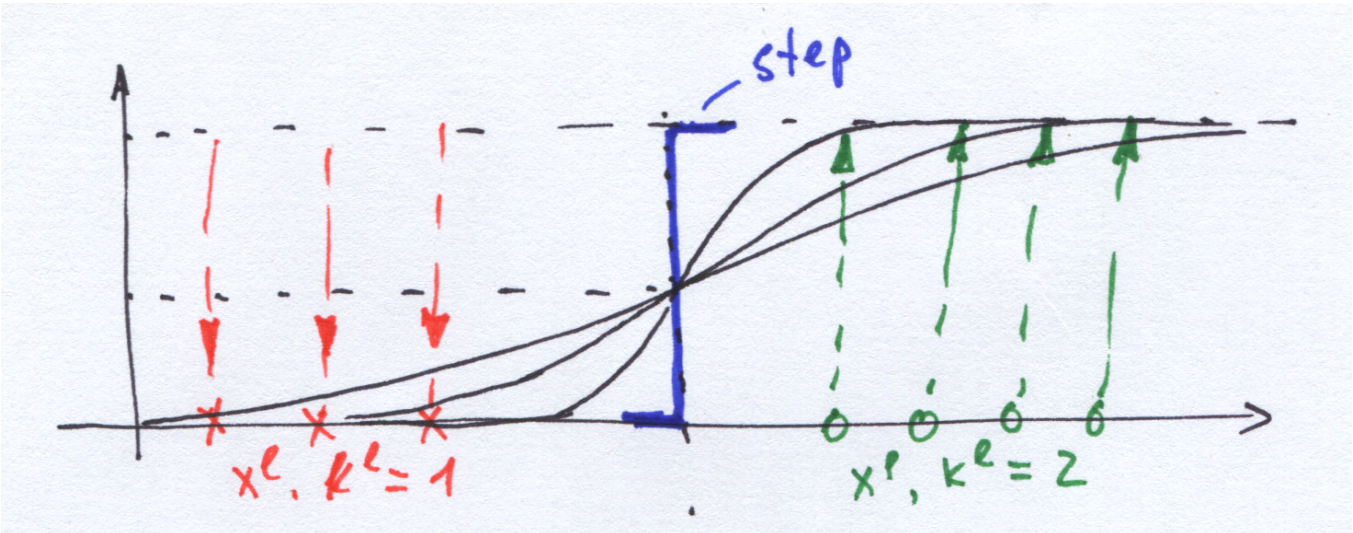
1. We learn the joint p.d. using the "usual" Maximum Likelihood (see the previous lecture). Then we derive the parameters of the posterior w and b from the learned p, σ, μ
2. We learn w and b by maximizing the Conditional Likelihood

Question: do these two schemes lead to the same parameters?

Generative vs. discriminative

Answer: “no” in general.

Counterexample: let there exist parameter values w and b for which $k^l = 2 \Leftrightarrow p(k{=}2|x^l) > p(k{=}1|x^l)$ for all l



Conditional Likelihood maximizes $p(1|x^l)$ for l with $k^l = 1$ and $p(2|x^l)$ for l with $k^l = 2$. The sigmoid-function becomes a step-function, which corresponds to $\sigma \to 0$ or $|\mu^2 - \mu^1| \to \infty$

Generative vs. discriminative

No fully unsupervised learning in the discriminative case :-(

For an incomplete training set $L = (x^1, x^2 \dots x^l)$

$$\ln p(L;\theta) = \sum_l \ln \sum_k p(x^l, k; \theta) =$$

$$= \sum_l \ln \sum_k \Big[p(x^l) \cdot p(k|x^l;\theta) \Big] = \sum_l \ln p(x^l)$$

$\rightarrow$ does not depend on the parameter at all.

Discriminant functions

- Let a parameterized family of p.d.-s be given.
- If the loss-function is fixed, each p.d. leads to a classifier
- The final goal is the classification (applying the classifier)

Generative approach:

1. Learn the parameters of the p.d. (e.g. ML)
2. Derive the corresponding classifier (e.g. Bayes)
3. Apply the classifier for test data

Discriminative (non-statistical) approach:

1. Learn the unknown parameters of the classifier directly
2. Apply the classifier for test data

If the family of classifiers is "well parameterized", it is not necessary to consider the underlying p.d. at all !!!

Linear discriminant functions

As before: two Gaussians of the same variance, known prior
Now: let the loss function be δ so the decision strategy is MAP

Remember the posterior:

$$p(k{=}1|x) = \frac{1}{1 + \exp\Big(\langle x, w\rangle + b\Big)}$$

$\rightarrow$ the classifier is given by $\langle x, w\rangle \lessgtr b$

It defines a **hyperplane** orthogonal to w that is shifted from the origin by $b/||w||$

Note: for the classifier it does not matter, how strong (step-like) is the underlying sigmoid-function $\rightarrow$ the variance σ is irrelevant $\rightarrow$ the classifier has even less free parameters then the corresponding posterior

Empirical Risk

How to find a good classifier ?

Bayesian risk:

$$R_b(e) = \sum_x \sum_k p(k,x) C\Big(e(x), k\Big) \rightarrow \min_e$$

But now it can not be computed because there is no p.d. !!!

We have only the training set $L = \Big((x^l, k^l) \ldots\Big)$

The **Bayesian** risk is replaced by the **Empirical** one – average loss over the training set instead of over the whole space:

$$R_e(e) = \sum_l C\Big(e(x^l), k^l\Big) \rightarrow \min_{e \in \mathcal{E}}$$

with a predefined classifier family $\mathcal{E}$.

Vapnik-Chervonenkis Dimension

Is the learning good (enough) ?

A reasonable measure would be the reached Bayesian risk. However, it can not be computed since there is no probability model. However, one can compute the Empirical risk.

$\rightarrow$ The question: how fast (and whether at all) does the Empirical risk converges to the Bayesian one with the increase of the training set N ?

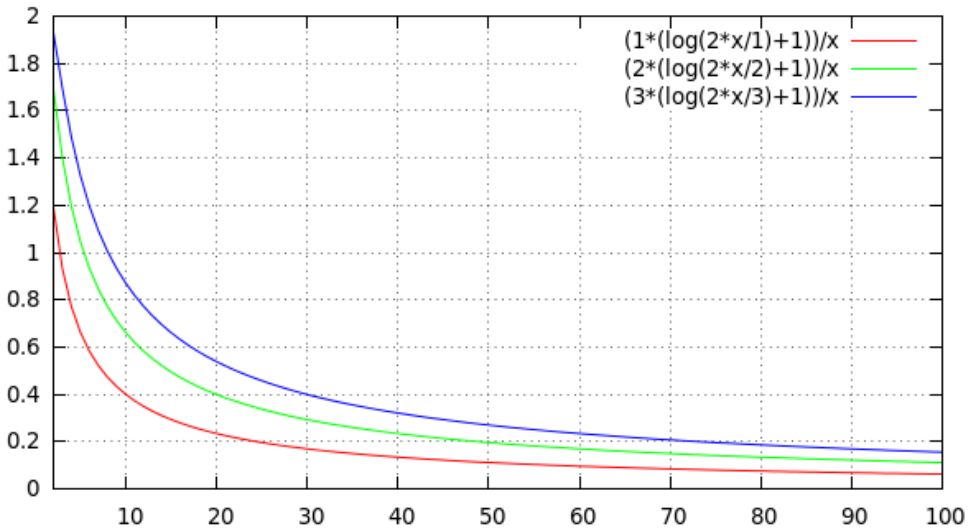
Upper bound for the difference (Vapnik, Chervonenkis, 1968):

$$P\left\{|R_b - R_e| < \sqrt{\frac{h\Big(\log(2N/h)+1\Big) - \log(\delta/4)}{N}}\right\} > 1-\delta$$

"The probability (over all training sets) that the considered difference is less then something is greater as something".

Vapnik-Chervonenkis Dimension

$$P\left\{|R_b - R_e| < \sqrt{\frac{h\Big(\log(2N/h)+1\Big) - \log(\delta/4)}{N}}\right\} > 1-\delta$$



The convergence speed depends on a constant h, which is called Vapnik-Chervonenkis Dimension. It reflects the "power" of the classifier family. The greater VC the worse the **generalization capabilities** of the classifier family.

Vapnik-Chervonenkis Dimension

A constructive definition:

A classifier family **shatters** the set of data points if, for **all** classifications of these points, there **exists** a classifier such that the model makes no errors when evaluating that set of data points.

The VC-Dimension of the family is the **maximal** number of points that **can** be arranged so that the family shatters them.

Alternative: The VC-Dimension is the **smallest** number of data points so that for **any** arrangement there **exists** a classification that **can not** be re-produced by the family.

Example: for linear classifiers in $\mathbb{R}^n$ the VC-dimension is $VC = n + 1$ (see the board).

Vapnik-Chervonenkis Dimension

The VC-dimension is often related to the number of free parameters (but not always, example – sinus, one free parameter, infinite VC)

The lower is VC the more robust is the family of classifiers.

Dilemma: complex data $\rightarrow$ complex classifiers (to reach **good recognition** rate) $\rightarrow$ many free parameters (high VC) $\rightarrow$ **bad generalization** capabilities.

Overfitting:
the classifier specializes to a particular training set.

Classifiers vs. generative models

Families of classifiers are usually "simpler" compared to the corresponding families of probability distributions (lower dimensions, less restricted etc.)

Often it is not necessary to care about the model consistency (such as e.g. normalization) $\rightarrow$ algorithms become simpler.

It is possible to use more complex decision strategies, i.e. to reach better recognition results.

However:

Large classified training sets are usually necessary, unsupervised learning is not possible at all.

Worse generalization capabilities, overfitting.

Conclusion – a "hierarchy of abstraction"

1. **Generative models** (joint probability distributions) represent the **entire** "world". At the learning stage (ML) the **probability** of the training set is maximized, no loss function.

2. **Discriminative models** represent posterior probability distributions, i.e. only what is needed for recognition. At the learning stage (ML) the **conditional likelihood** is maximized, no loss function.

3. **Discriminant functions**: no probability distribution, decision strategy is learned directly, the **Empirical risk** is minimized.