Statistical Machine Learning (BE4M33SSU)
Lecture 8: Generative learning, Maximum Likelihood Estimator

Czech Technical University in Prague

- When do we need generative learning?
- Parametric distribution families
- Maximum Likelihood Estimator and its properties

Reminder: discriminative learning

Goal: train a classifier $y = h(x)$ for an unknown distribution $p(x, y)$ for features $x \in \mathcal{X}$ and classes $y \in \mathcal{Y}$

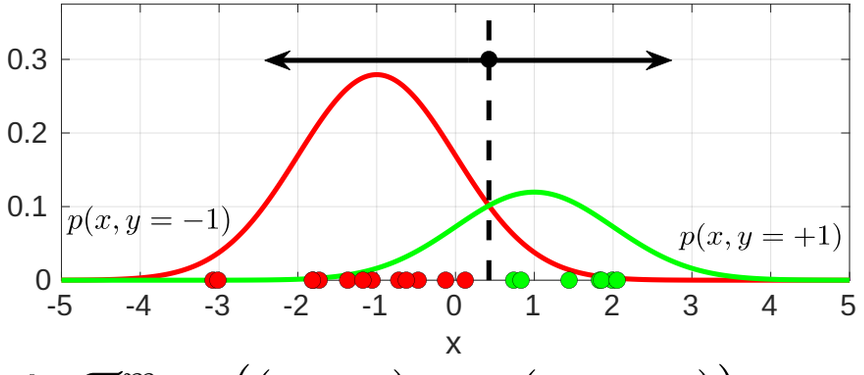
Discriminative learning:

- define a hypothesis space $\mathcal{H}$ of predictors $h\colon \mathcal{X} \to \mathcal{Y}$ and fix a loss $\ell(y, y')$
- given a training set $\mathcal{T}^m$, learn $h_m\colon \mathcal{X} \to \mathcal{Y}$ by empirical risk minimisation.

Example 1 (Gaussian discriminative analysis)**.** Assume we know: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{-1, +1\}$

$$p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}$$

with unknown $p(y = 1)$, $\mu_+ > \mu_-$ and σ.



The loss is $\ell(y, y') = [\![y' \neq y]\!]$ and the training set is $\mathcal{T}^m = \big((x_1, y_1), \ldots, (x_m, y_m)\big)$.

- The Bayes optimal predictor for each such model is $\operatorname{sign}(x - \theta)$ with some threshold θ, thus $\mathcal{H}$ has VC-dimension $d = 1$.
- We apply empirical risk minimisation and want to bound the estimation error $R(h_{\mathcal{H}}) - R(h_m)$ by $\epsilon = 0.03$ with probability 0.95 over training sets $\mathcal{T}^m$.

From theory: $\mathcal{H}$ satisfies the ULLN $\Rightarrow$ ERM is a successful PAC-learner $\Rightarrow$ we need $m > 10^5$ training examples.

Generative learning (Setup)

Generative learning: Use prior knowledge to restrict the search to a parametric family of distributions $p_\theta(x,y)$, $\theta \in \Theta$. Learning algorithm:

1. Given training data $\mathcal{T}^m$, estimate the unknown parameter $\theta_m = e(\mathcal{T}^m)$ e.g. using the maximum likelihood estimator.
2. Consider $p_{\theta_m}(x,y)$ as the true model. Predict hidden states by its Bayes optimal predictor

$$h(x) = \operatorname*{arg\,min}_{y \in \mathcal{Y}} \sum_{y' \in \mathcal{Y}} p_{\theta_m}(y' \,|\, x)\, \ell(y', y).$$

Example 1 (cont.)**.** Given $\mathcal{T}^m$, the estimates of the model parameters are

$$p(y=1) = \frac{m_+}{m} \quad \mu_+ = \frac{1}{m_+} \sum_i x_i [\![y_i = 1]\!] \quad \mu_- = \frac{1}{m_-} \sum_i x_i [\![y_i = -1]\!]$$

and

$$\sigma^2 = \frac{1}{m} \sum_i \Big(x_i - \mu_+ [\![y_i = 1]\!] - \mu_- [\![y_i = -1]\!] \Big)^2,$$

where m_+ denotes the number of training examples with class $y_i = 1$. The predictor is

$$h(x) = \operatorname{sign}\big[p(x, y=1) - p(x, y=-1) \big] = \ldots = \operatorname{sign}(x - \theta),$$

where θ depends on the estimated μ_+, μ_-, σ, $p(y=1)$ and $p(y=-1)$.

When do we need generative learning?

We can not prove that this leads to a successful PAC-learner.

When and why shall we use generative learning?

- if we need the uncertainty of the prediction $h_m(x)$,
- if we want to detect outliers when predicting,
- for semi-supervised learning, i.e. when only a part of the training data is annotated,
- if the statistical relation between x and y depends on some *latent variables* z, e.g. $p(x, y, z) = p(x \mid z, y)p(z)p(y)$ and they are not accessible for training,
- if we want to learn models that can generate realistic data x.

Parametric distribution families

A *parametric distribution family* is a set of distributions for a r.v. X which are specified by parameter values.

Example 2. The family of multivariate normal distributions $\mathcal{N}(\mu, V)$ on $\mathbb{R}^n$

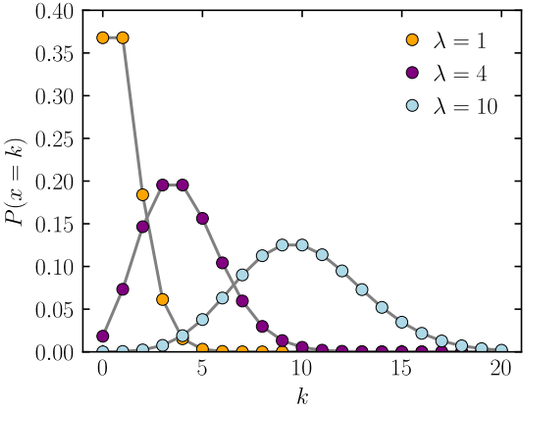
$$p_{\mu,V}(x) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\Big[-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu)\Big]$$

parametrised by the vector $\mu \in \mathbb{R}^n$ and a positive (semi) definite $n \times n$ matrix V.

Example 3. The family of Poisson distributions on $x \in \mathbb{N}$ with probability mass

$$p(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

parametrised by $\lambda \in \mathbb{R}_+$. Notice that $\lambda = \mathbb{E}[X] = \mathbb{V}[X]$.



Both families are examples of a broad class of distribution families – *exponential families*.

Parametric distribution families

Definition 1. A family of distributions for a random variable $x \in \mathcal{X}$ is an *exponential family* if its probability density / probability mass has the form

$$p_\theta(x) = h(x) \exp\big[\langle \phi(x), \theta \rangle - A(\theta)\big],$$

where

$\phi(x) \in \mathbb{R}^n$ is the sufficient statistics,

$\theta \in \mathbb{R}^n$ is the (natural) parameter,

$h(x)$ is the base measure and

$A(\theta)$ is the cumulant function defined by

$$A(\theta) = \log \int_{\mathbb{R}^n} h(x) \exp\big[\langle \phi(x), \theta \rangle\big] \, d\nu(x)$$

Notes:

- The cumulant function is essentially the logarithm of the normalisation constant.
- The statistic $\phi(x)$ is called *sufficient* because when estimating the parameter θ from a training set $\mathcal{T}$, all we need to know from it is $\mathbb{E}_{\mathcal{T}}[\phi(x)]$.

Parametric distribution families

Example 4. Consider the family of Bernoulli distributions for $x \in \{0, 1\}$ with $p(x) = \beta^x (1-\beta)^{1-x}$ parametrised by $\beta \in (0, 1)$. It can be written as

$$p(x) = \exp[\langle \phi(x), \theta \rangle - A(\theta)]$$

with $\phi(x) = x$, $\theta = \log \frac{\beta}{1-\beta}$ and $A(\theta) = \ldots$.

Example 5. Consider the family of univariate normal distributions with unit variance and mean μ for $x \in \mathbb{R}$. Its density is given by

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$$

and can be written as

$$p(x) = h(x) \exp\big[\langle \phi(x), \theta \rangle - A(\theta)\big]$$

with $h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $\phi(x) = x$, $\theta = \mu$ and $A(\theta) = \frac{\theta^2}{2}$.

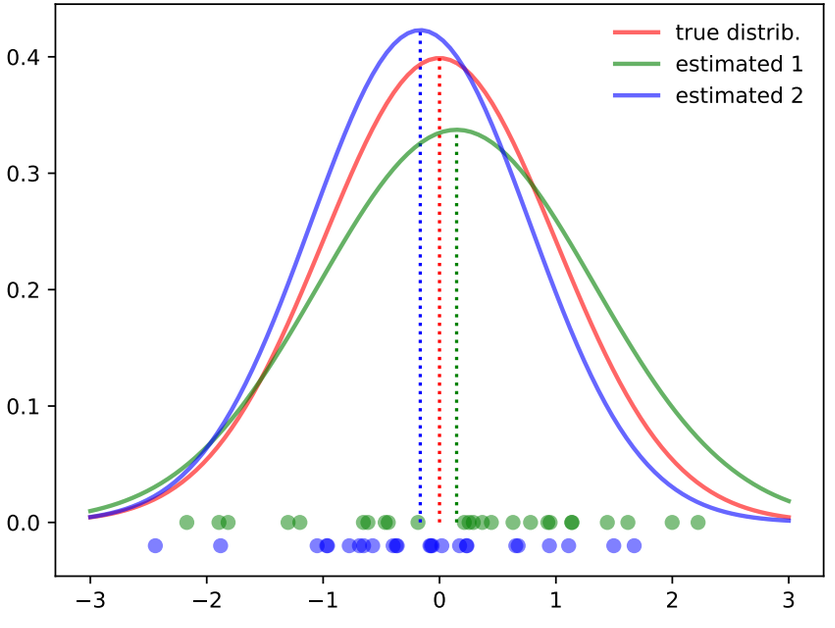
Parameter estimation

Given: a parametric family of distributions $p_\theta(x)$, $\theta \in \Theta$ and an i.i.d. training set $\mathcal{T}^m = \{x^j \in \mathcal{X} \mid j = 1, \ldots, m\}$ generated from $p_{\theta^*}(x)$ with unknown θ^*.

Estimator: a mapping $\theta_m = e(\mathcal{T}^m)$, which maps training sets to parameters, i.e. $e\colon \mathcal{T}^m \mapsto \theta_m \in \Theta$

Example 6. Estimating parameters of a normal distribution

- red: true distribution $\mathcal{N}(0, 1)$
- blue and green: sample two i.i.d. training sets from it and estimate parameters.



Desired properties of an estimator:

- the estimator is unbiased i.e. $\mathbb{E}_{\mathcal{T}^m \sim \theta^*}\left[e(\mathcal{T}^m)\right] = \theta^*$
- the estimator has small variance $\mathbb{V}_{\mathcal{T}^m \sim \theta^*}\left[e(\mathcal{T}^m)\right] \to 0$ for $m \to \infty$
- the estimator is consistent $\mathbb{P}_{\theta^*}\Big(|e(\mathcal{T}^m) - \theta^*| \geqslant \epsilon\Big) \to 0$ for $m \to \infty$

Maximum Likelihood estimator

Define the log-likelihood to obtain the given i.i.d. training data $\mathcal{T}^m$ from the distribution with parameter $\theta \in \Theta$

$$L_{\mathcal{T}^m}(\theta) = \frac{1}{m} \log \mathbb{P}_\theta(\mathcal{T}^m) = \frac{1}{m} \sum_{x \in \mathcal{T}^m} \log p_\theta(x)$$

Notice: we normalise the log-likelihood by the sample size to make it comparable for different sample sizes.

The **Maximum Likelihood estimator** predicts the parameter θ_m that maximises the (log-) likelihood of the training data

$$\theta_m = e_{ML}(\mathcal{T}^m) \in \arg\max_{\theta \in \Theta} L_{\mathcal{T}^m}(\theta) = \arg\max_{\theta \in \Theta} \frac{1}{m} \sum_{x \in \mathcal{T}^m} \log p_\theta(x)$$

Maximum Likelihood estimator

Example 7 (MLE for exponential families)**.** Consider the parametric family

$$p_\theta(x) = \exp[\langle \phi(x), \theta \rangle - A(\theta)]$$

with sufficient statistic $\phi(x) \in \mathbb{R}^n$ and natural parameter $\theta \in \mathbb{R}^n$. We are given an i.i.d. training set $\mathcal{T}^m$ and want to estimate θ by the MLE. The log-likelihood $L_{\mathcal{T}^m}(\theta)$ is a concave function of θ and has only global maxima (see seminar). We compute its derivative and set it to zero.

$$\begin{aligned}\nabla L_{\mathcal{T}^m}(\theta) &= \nabla \frac{1}{m} \sum_{x \in \mathcal{T}^m} \big[\langle \phi(x), \theta \rangle - A(\theta)\big] = \frac{1}{m} \sum_{x \in \mathcal{T}^m} \phi(x) - \nabla \log \sum_{x \in \mathcal{X}} e^{\langle \phi(x), \theta \rangle} \\ &= \mathbb{E}_{\mathcal{T}^m}[\phi(x)] - \mathbb{E}_\theta[\phi(x)] = 0\end{aligned}$$

The maximum likelihood estimator picks θ so that the expectation of $\phi(x)$ under the model coincides with its empirical expectation on the training data.

Usually, we cannot compute this estimator in closed form, but we can use e.g. gradient ascent to find the maximum.

Kullback-Leibler divergence

To analyse properties of the ML-estimator, we will need the following similarity measure for distributions.

Kullback-Leibler divergence: for distributions $p(x)$, $q(x)$ defined on $\mathcal{X}$

$$D_{KL}(q(x) \,\|\, p(x)) = \sum_{x \in \mathcal{X}} q(x) \log \frac{q(x)}{p(x)}$$

D_{KL} is non-negative, i.e. $D_{KL}(q(x) \,\|\, p(x)) \geqslant 0$ with equality iff $p(x) = q(x)\ \forall x \in \mathcal{X}$. This follows from strict concavity of the function $\log(x)$

$$-D_{KL}(q \,\|\, p) = \sum_{x \in \mathcal{X}} q(x) \log \frac{p(x)}{q(x)} \leqslant \sum_{x \in \mathcal{X}} q(x) \left[\frac{p(x)}{q(x)} - 1 \right] = 0$$

- it is not symmetric, i.e. $D_{KL}(q(x) \,\|\, p(x)) \neq D_{KL}(p(x) \,\|\, q(x))$.
- it is undefined if $\exists x \colon q(x) > 0$ and $p(x) = 0$.
- D_{KL} can be generalised for continuous distributions and is invariant under coordinate transforms.

Properties of the ML estimator

(1) Is the Maximum Likelihood estimator unbiased?

No, it is not unbiased in general.

(2) What conditions ensure MLE consistency, i.e.

$$\mathbb{P}_{\theta^*}\big(|\theta^* - e_{ML}(\mathcal{T}^m)| > \epsilon\big) \xrightarrow{m\to\infty} 0,$$

where probability is w.r.t. $\mathcal{T}^m \sim p_{\theta^*}(x)$?

To answer this question, we first notice that the ML estimator is equivalent to

$$R_m(\theta, \theta^*) = \frac{1}{m} \sum_{x \in \mathcal{T}^m} \log \frac{p_{\theta^*}(x)}{p_\theta(x)} = L_{\mathcal{T}^m}(\theta^*) - L_{\mathcal{T}^m}(\theta) \to \min_\theta$$

where $R_m(\theta, \theta^*)$ is a sample estimate of the KL divergence $D_{KL}(p_{\theta^*}(x) \,\|\, p_\theta(x)) =: R(\theta, \theta^*)$

Properties of the ML estimator

The ML estimator is consistent if the following two properties hold:

Condition 1 Identifiability: If $\theta_1 \neq \theta_2$ then $p_{\theta_1}(x) \neq p_{\theta_2}(x)$. Moreover, we require

$$\inf_{\theta:\ \|\theta-\theta_*\|>\epsilon} D_{KL}(p_{\theta^*}(x) \,\|\, p_\theta(x)) > 0$$

Condition 2 Uniform LLN

$$\mathbb{P}\Big(\sup_\theta |R_m(\theta,\theta^*) - R(\theta,\theta^*)| > \epsilon\Big) \to 0 \text{ for } m \to \infty.$$

Example 8. For an exponential family $p_\theta(x) = \exp[\langle \phi(x), \theta\rangle - A(\theta)]$ we have

$$R_m(\theta,\theta^*) - R(\theta,\theta^*) = \mathbb{E}_{\mathcal{T}^m}\big[\langle \phi(x), \theta-\theta^*\rangle\big] - \mathbb{E}_{\theta^*}\big[\langle \phi(x), \theta-\theta^*\rangle\big],$$

and it is then usually easy to specify conditions under which the ULNN holds.

Properties of the ML estimator

What can we say about the variance of the ML estimator, i.e. $\mathbb{V}_{\mathcal{T}^m \sim \theta^*}\big[e_{ML}(\mathcal{T}^m)\big]$?

The asymptotic variance of the ML estimator is, in a certain sense, the smallest possible!

To make this precise, we need the notion of *Fisher information*

$$I(\theta) = \int \Big[\frac{d}{d\theta} \log p_\theta(x)\Big]^2 p_\theta(x)\, dx = \mathbb{E}_\theta \Big[\frac{d}{d\theta} \log p_\theta(x)\Big]^2$$

It is easy to show that $\mathbb{E}_\theta\big[\frac{d}{d\theta} \log p_\theta(x)\big] = 0$ (see seminar). Therefore, the Fisher information is the variance of the random variable $\frac{d}{d\theta} \log p_\theta(x)$.

Now, we have the following two statements about the variance of estimators

- The asymptotic distribution of the ML estimator is:

$$e_{ML}(\mathcal{T}^m) \sim \mathcal{N}\Big(\theta^*, \frac{1}{mI(\theta^*)}\Big) \quad \text{for } m \to \infty$$

- If e is an arbitrary unbiased estimator, then its variance can not be smaller, i.e.

$$\mathbb{V}_{\mathcal{T}^m \sim \theta^*}\big[e(\mathcal{T}^m)\big] \geqslant \frac{1}{mI(\theta^*)}$$

Properties of the ML estimator

Summary:

- ML estimator can be biased,
- ML estimator is consistent under mild conditions,
- ML estimator has asymptotically optimal variance.

Remark 1 (model misspecification)**.** In machine learning we usually do not believe that the model class $p_\theta(x)$, $\theta \in \Theta$ is chosen correctly, i.e. such that it contains the unknown data distribution. It is rather believed to be an useful idealisation. Let $q(x)$ denote the true but unknown data distribution and let $\theta_m = e_{ML}(\mathcal{T}^m)$ be the ML estimate for a given training set. Using the same arguments as above, we see for the limit $m \to \infty$

$$D_{KL}(q(x) \,\|\, p_{\theta_m}(x)) \leqslant D_{KL}(q(x) \,\|\, p_\theta(x)) \quad \forall \theta \in \Theta$$

This means, that the MLE gives the *KL-projection* of q on our model class.