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

1. When do we need generative learning?

Discriminative learning: $p(x, y)$ unknown

- define a hypothesis class $\mathcal{H}$ 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.

Cases when this is not sufficient:

- we need the uncertainty of the prediction $h_m(x)$
- semi-supervised learning, i.e. only a part of the training data is annotated
- the statistical relation between x and y depends on some *latent variables* z, e.g. $p(x, y, z) = p(x \mid z, y)p(z)p(y)$, but we never see z in the training data.
- we want to learn models that can generate realistic data x

1. When do we need generative learning?

Generative learning:

- prior knowledge/assumption: define a parametric family of distributions $p_\theta(x, y)$, $\theta \in \Theta$
- given training data $\mathcal{T}^m$, estimate the unknown parameter $\theta_m = e(\mathcal{T}^m)$.
- Then predict hidden states by

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

- the uncertainty of the prediction can be obtained from $p_{\theta_m}(y \,|\, x)$,
- data can be generated from $p_{\theta_m}(x \,|\, y)$.
- semi-supervised learning possible e.g. by Expectation Maximisation algorithm

2. Parametric distribution families

Parametric distribution family: A set of distributions for a r.v. X with common structure and specified by parameter values.

Example 1. 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 2. 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 $\lambda = \mathbb{E}[X] = \mathbb{V}[X]$.

2. Parametric distribution families

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

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

2. Parametric distribution families

Kullback-Leibler divergence: similarity measure for distributions, defined by

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

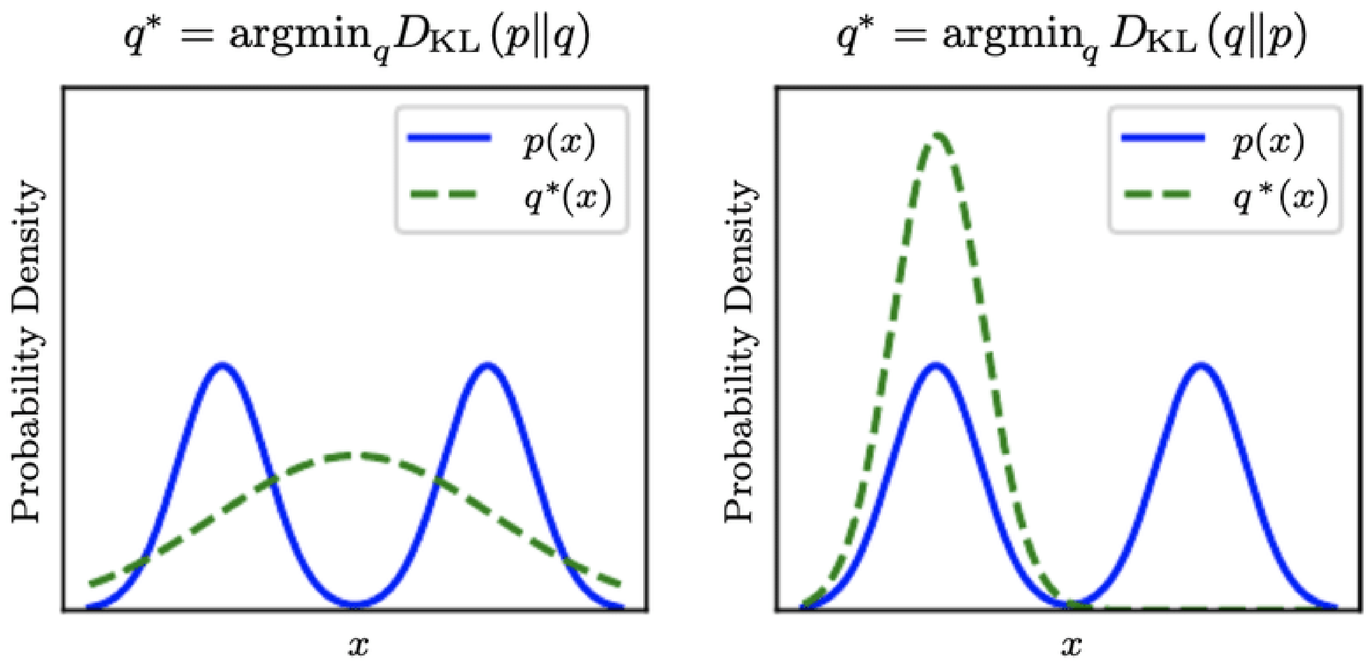
D_{KL} is non-negative, i.e. $D_{KL}(q(x) \parallel p(x)) \geqslant 0$ with equality iff $p(x) = q(x)\ \forall x \in \mathcal{X}$. This follows from strict concavity of the function $\log(x)$

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

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

2. Parametric distribution families

Example 3. Approximate a mixture of two Gaussians $p(x)$ by a single Gaussian $q(x)$ w.r.t. KL-divergence. Difference between forward and reverse KL-divergence.



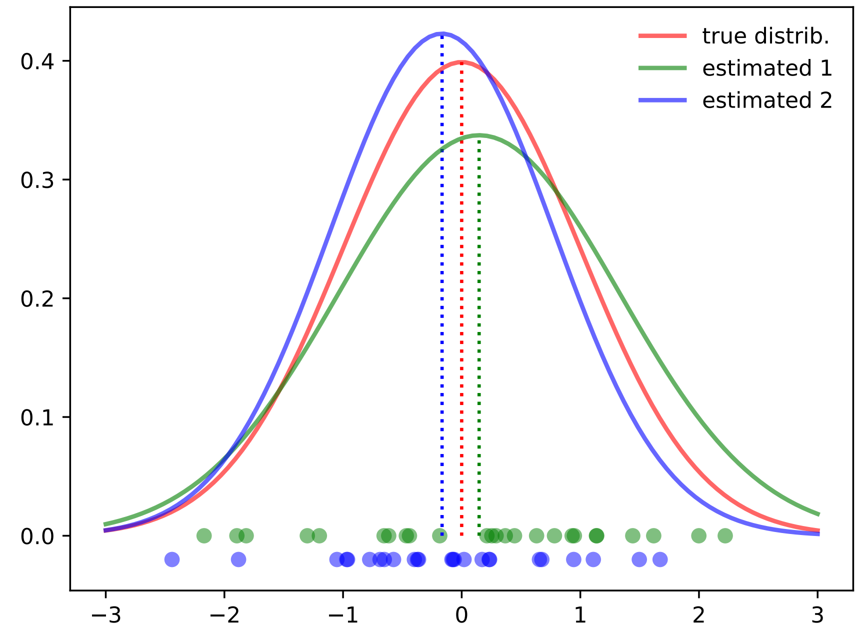
3. Parameter estimation

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: 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:

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

3. 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** is defined by

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

i.e. the estimate θ_m is a maximiser of the log-likelihood.

Is the Maximum Likelihood estimator unbiased?

No, it is not unbiased in general.

3. Maximum Likelihood estimator

What conditions ensure MLE consistency, i.e.

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

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

The ML estimator is consistent if the following properties hold:

- the parameter set $\Theta \in \mathbb{R}$ is an open interval,
- the density is strictly positive, i.e. $p_\theta(x) > 0$, and is differentiable in θ for all x,
- the equation

  $$\frac{d}{d\theta} L_{\mathcal{T}^m}(\theta) = \frac{d}{d\theta}\Big[\frac{1}{m}\sum_{x\in\mathcal{X}} \log p_\theta(x)\Big] = 0$$

  has exactly one solution which corresponds to a maximum of $L_{\mathcal{T}^m}(\theta)$. This holds for each m and each training set $\mathcal{T}^m$.

This can be generalised to the case of many parameters $\Theta \in \mathbb{R}^n$.

3. Maximum Likelihood 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*

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

Under some regularity conditions, we have

$$\int \frac{d}{d\theta} p_\theta(x)\, dx = 0 \text{ and } \int \frac{d^2}{d\theta^2} p_\theta(x)\, dx = 0.$$

Then we have the following equivalent definitions of Fisher information:

$$I(\theta) = \mathbb{V}_\theta \Big[\frac{d}{d\theta} \log p_\theta(x)\Big] \text{ and } I(\theta) = -\mathbb{E}_\theta \Big[\frac{d^2}{d\theta^2} \log p_\theta(x)\Big]$$

3. Maximum Likelihood estimator

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 unbiased estimator, then its variance can not be smaller, i.e.

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

Summary:

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

3. Maximum Likelihood estimator

Example 4 (MLE for an exponential family)**.** Let us consider an exponential family

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

and the ML estimator for an i.i.d. training set $\mathcal{T}^m = \{x_i \mid i = 1 \ldots, m\}$. Its log-likelihood is

$$L_{\mathcal{T}^m}(\theta) = \frac{1}{m}\sum_{x\in\mathcal{T}^m} \log p_\theta(x) = \frac{1}{m}\sum_{x\in\mathcal{T}^m} \langle \phi(x), \theta\rangle - A(\theta) = \langle \psi, \theta\rangle - A(\theta),$$

where we denoted $\psi = \mathbb{E}_{\mathcal{T}^m}[\phi(x)]$.

- sufficient statistics: we need to now $\mathbb{E}_{\mathcal{T}^m}[\phi(x)]$ only.
- The function $A(\theta)$ is convex and has gradient $\nabla A(\theta) = \mathbb{E}_\theta[\phi]$ (see seminar).
- $L_{\mathcal{T}^m}(\theta)$ is concave. Hence any critical point θ with $\nabla L_{\mathcal{T}^m}(\theta) = 0$ is a global maximum.
- Maximisers θ^* are given by the equation $\mathbb{E}_{\mathcal{T}^m}[\phi] = \mathbb{E}_{\theta^*}[\phi]$.
- The Fisher information for the family is given by the variance of the sufficient statistics

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