OI-OPPA. European Social Fund
Prague & EU: We invests in your future.

Machine Learning
Lecture 11

Thomas Hofmann

June 2, 2005

Statistical Estimation

- General setting:
  - **random variable** X = a variable representing a random event
  - **sample space** $\mathcal{X}$ = space of possible outcomes
  - **realization** $\mathbf{x}$ = observed or hypothetical outcome
  - set of **probability distributions** $\mathcal{P}$ over $\mathcal{X}$ parameterized by some **parameter vector** θ, $p(\cdot;\theta) \in \mathcal{P}$.
  - E.g. $p(\cdot;\theta) \geq$ is a probabiliy density function $\int_{\mathcal{X}} p(\mathbf{x};\theta) d\mathbf{x} = 1$

- **Statistical estimation**: given an observation or a set of observations, infer an **optimal parameter** θ

Maximum Likelihood Estimation

- Use likelihood as criterion to rate different hypotheses (θ).

- More convenient tu use so-called **log-likelihood function**

$$\mathcal{L}(\theta; \mathbf{x}) = \log p(\mathbf{x}; \theta)$$

- This means, a parameter θ is preferred over some $\bar{\theta}$, if the observed data is more likely under θ than $\bar{\theta}$.

- **Maximum Likelihood Estimation**

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta; \mathbf{x}) = \arg\max_{\theta} \log p(\mathbf{x}; \theta)$$

- i.i.d. sample $\mathbf{x} = \mathbf{x}_1, \ldots, \mathbf{x}_n$: $\hat{\theta} = \arg\max_\theta \sum_{i=1}^n \log p(\mathbf{x}_i; \theta)$

MLE: Gaussian Case

- Example: Gaussian distribution, $\mathcal{X} = \mathbb{R}$, $\theta = (\mu, \sigma)'$, probability density

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

- Maximum likelihood estimates

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \end{aligned}$$

MLE: Multivariate Normal Distribution

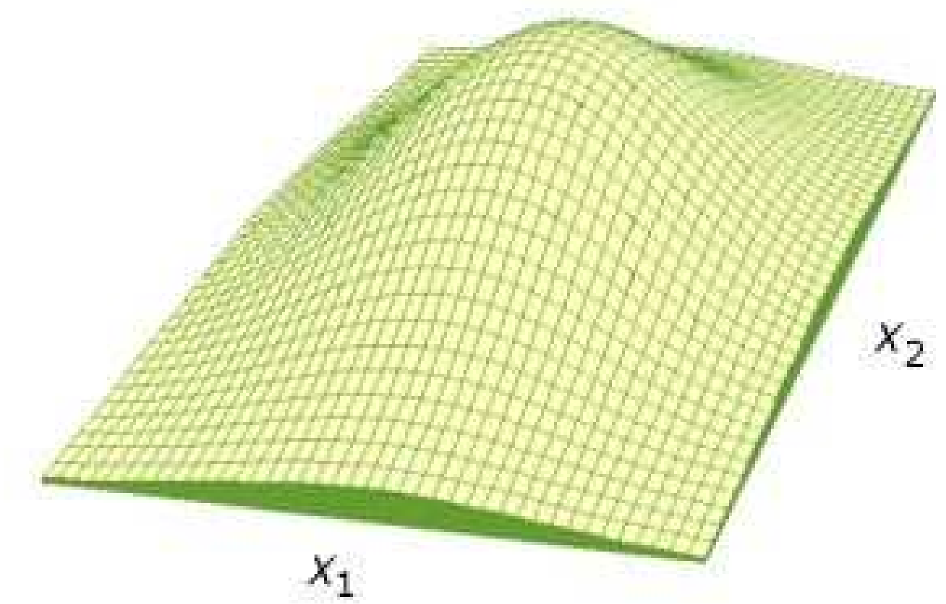
- Multivariate normal

$$p(\mathbf{x};\mu,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

- MLE

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \\ \hat{\mathbf{\Sigma}} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\hat{\mu})(\mathbf{x}_i-\hat{\mu})' \end{aligned}$$

MLE: Multivariate Normal Distribution



Mixture Models (1)

- Statistical classification: assume that the observed patterns $\mathbf{x}_1, \ldots, \mathbf{x}_n$ belong to a certain number of K **classes** $c_1, \ldots, c_K$.

- Assume further that we do not observe these classes, but rather a mixture of patterns from different classes.

- For each class we assume that patterns are distributed according to a **class-conditional distribution** $p_k(\mathbf{x}; \theta_k)$ parameterized by θ_k, $p_k(\mathbf{x}; \theta) = p(\mathbf{x}|\mathbf{C} = c_k; \theta_k)$. Denote $\theta = (\theta_1, \ldots, \theta_K)'$.

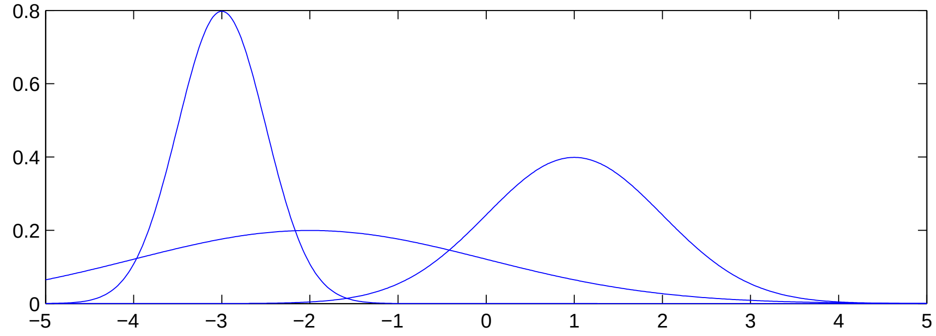
- These assumptions lead to a **mixture model**

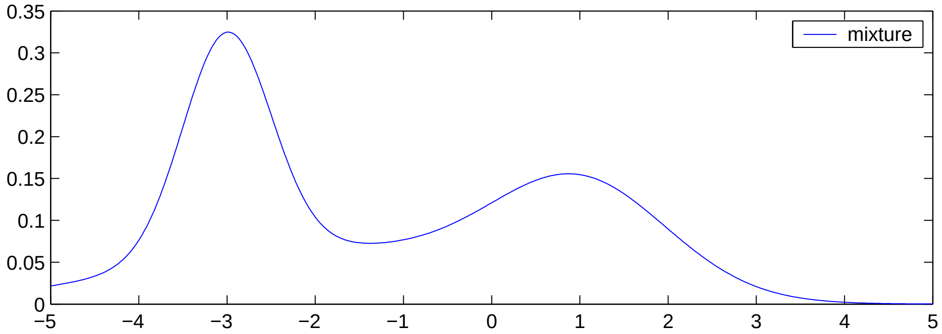
$$p(\mathbf{x}; \pi, \theta) = \sum_{k=1}^{K} \pi_k \, p_k(\mathbf{x}; \theta_k)$$

where π_k is the prior probability of class c_k (mixing proportions).

Mixture Models (2)

- Notice that $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$.
- A simple example of a density consisting of a mixture of three Gaussians





Why Mixture Models?

- Mixture models are more powerful than the component models used for the class-conditional distribution
- Mixture models can capture multimodality and offer a systematic way to define complex statistical models based on simpler ones.
- Mixture models can also be utilized to "unmix" the data, i.e. to assign patterns to the unobserved classes (**data clustering**)
- Bayes rule: **posterior probabilities**

$$P(c_k|\mathbf{x};\pi,\theta) = \frac{\pi_k \cdot p_k(\mathbf{x};\theta_k)}{\sum_{l=1}^{K} \pi_l \cdot p_l(\mathbf{x};\theta_l)}$$

MLE in Mixtures: Complete Data Log-Likelihood

- Key question: how to fit the parameters π, θ of a mixture model
- **Expectation Maximization (EM) algorithm**
- Introduce unobserved cluster membership variables $z_{ik} \in \{0, 1\}$
  - $z_{ik} = 1$ denotes the fact that data point $\mathbf{x}_i$ has been generated from the k-th component or class
  - $\sum_{k=1}^{K} z_{ik} = 1$ for all $i = 1, \ldots, n$
- If membership variables were observed, then one could define the so-called **complete data log-likelihood**,

$$\mathcal{L}_c(\pi, \theta; \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left[\log p_k(\mathbf{x}_i; \theta_k) + \log \pi_k \right]$$

MLE in Mixtures: Observed Data Log-Likelihood

- Since class membership variables z are not observed, we only have access to the **observed data log-likelihood**

$$\mathcal{L}(\pi, \theta; \mathbf{x}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i; \theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i; \theta_k),$$

- Problem: direct maximization is difficult (logarithm of a sum effectively introduces complicated couplings)

Statistical Models with Unobserved Variables

- Imagine we would have some estimate of what the unobserved variables could be:

$$Q_{ik} = \mathsf{Pr}(z_{ik} = 1) = \text{probability that } \mathbf{x}_i \text{ belongs to cluster } c_k$$

- Try to maximize the expected complete data log-likelihood

$$\mathbf{E}_Q\left[\mathcal{L}_c(\pi, \theta; \mathbf{x}, \mathbf{z})\right] = \sum_{i=1}^{n} \sum_{k=1}^{K} Q_{ik}\left[\log p_k(\mathbf{x}_i; \theta_k) + \log \pi_k\right] \, .$$

- Q is called a **variational distribution** (we don't know yet how to chose it appropriately)

Expected Complete Data Log-Likelihood

- Consider the following line of argument

$$\begin{aligned}
\mathcal{L}(\pi, \theta; \mathbf{x}) &= \sum_{i=1}^{n} \log p(\mathbf{x}_i; \pi, \theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i; \theta_k) \\
&= \sum_{i=1}^{n} \log \sum_{k=1}^{K} Q_{ik} \frac{\pi_k p_k(\mathbf{x}_i; \theta_k)}{Q_{ik}} \\
&\geq \sum_{i=1}^{n} \sum_{k=1}^{K} Q_{ik} \log \frac{\pi_k p_k(\mathbf{x}_i; \theta_k)}{Q_{ik}} = L(\pi, \theta, Q; \mathbf{x})
\end{aligned}$$

- Inequality follows from the concavity of the logarithm, or more specifically from **Jensen's inequality**.

Jensen's Inequality

- Jensen's inequality: for a convex function f and any probability mass function p

$$\mathbf{E}[f(\mathbf{x})] = \sum_{\mathbf{x}} p(\mathbf{x}) f(\mathbf{x}) \geq f\left(\sum_{\mathbf{x}} p(\mathbf{x})\mathbf{x}\right) = f(\mathbf{E}[\mathbf{x}])$$

- Proof uses a simple inductive argument over the state space size.

Variational Upper Bound

- No matter what Q is, we will get a **lower bound** on the log-likelihood function.
- Instead of maximizing $\mathcal{L}$ directly, we can hence try to **maximize the (simpler) lower bound** $L(\theta, \pi, Q; \mathbf{x})$ w.r.t. the parameters θ and π.

Expectation Maximization Algorithm (1)

- Each choice of Q defines a different lower bound $L(\theta, \pi, Q; \mathbf{x})$
- Key idea: optimize lower bound also w.r.t. Q. Get **tightest lower bound** for a given estimate of θ.
- Alternation scheme, maximizes $L(\pi, \theta, Q; \mathbf{x})$ in every step.
  - **E-step**: $Q^{(t+1)} = \arg\max_Q L(\pi^{(t)}, \theta^{(t)}, Q; \mathbf{x})$
  - **M-step**: $(\pi^{(t+1)}, \theta^{(t+1)}) = \arg\max_{\pi,\theta} L(\theta, \pi, Q^{(t+1)}; \mathbf{x})$
- M-step optimizes a lower bound instead of the true likelihood function
- E-step adjusts the bound

Expectation Maximization Algorithm (2)

- What does that have to do with the function we referred to as expected complete data log-likelihood above?

$$
\begin{aligned}
L(\pi, \theta, Q; \mathbf{x}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} Q_{ik} \log \frac{\pi_k p_k(\mathbf{x}_i; \theta_k)}{Q_{ik}} \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} Q_{ik} \log \pi_k p_k(\mathbf{x}_i; \theta_k) - \sum_{i=1}^{n} \sum_{k=1}^{K} Q_{ik} \log Q_{ik} \\
&= \mathbf{E}_Q \left[\mathcal{L}_c(\pi, \theta; \mathbf{x}, \mathbf{z}) \right] - \sum_{i=1}^{n} \sum_{k=1}^{K} Q_{ik} \log Q_{ik}
\end{aligned}
$$

- Second term: entropy of Q (does not depend on π or θ)

- Maximizing $L(\pi, \theta, Q; \mathbf{x})$ is the same as maximizing the expected complete data log-likelihood.

Expectation Maximization Algorithm (3)

- How about the E-step?
- It is easy to find a general answer to how Q should be chosen.
- Posterior probability $Q^*_{ik} \equiv \Pr(z_{ik} = 1 | \mathbf{x}_i; \pi, \theta)$ maximizes $L(\pi, \theta, Q; \mathbf{x})$ for given π and θ.
- Proof: insert this choice for Q^* into $L(\pi, \theta, Q; \mathbf{x})$

$$
\begin{aligned}
L(\pi, \theta, Q^*; \mathbf{x}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} Q^*_{ik} \log \frac{\pi_k \, p_k(\mathbf{x}_i; \theta_k)}{Q^*_{ik}} \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} Q^*_{ik} \log p(\mathbf{x}_i; \pi, \theta) = \mathcal{L}(\pi, \theta; \mathbf{x})
\end{aligned}
$$

- Since $L(\pi, \theta, Q; \mathbf{x}) \leq \mathcal{L}(\pi, \theta; \mathbf{x})$ for all Q, equality is optimal.

Normal Mixture Model

- In the case of a mixture of multivariate normal distributions:
- M-step: differentiating expected complete data log-likelihood
- Mixing proportions $\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^n Q_{ik}$
- Normal model

$$\begin{aligned} \hat{\mu}_k &= \frac{\sum_{i=1}^n Q_{ik}\mathbf{x}_i}{\sum_{i=1}^n Q_{ik}} \\ \hat{\Sigma}_k &= \frac{\sum_{i=1}^n Q_{ik}(\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)'}{\sum_{i=1}^n Q_{ik}} . \end{aligned}$$

Normal Mixture Model (2)

- E-Step

$$Q_{ik} = \frac{\pi_k |\Sigma_k|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mu_k)\Sigma_k^{-1}(\mathbf{x}_i - \mu_k)\right]}{\sum_{l=1}^{K} \pi_l |\Sigma_l|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mu_l)\Sigma_l^{-1}(\mathbf{x}_i - \mu_l)\right]}$$

EM for Normal Mixture Model

initialize $\hat{\mu}_k$ at random

initialize $\hat{\Sigma}_k = \sigma^2 \mathbf{I}$, where σ^2 is the overall data variance

repeat

  for each data point $\mathbf{x}_i$ **do**

    for each component $k = 1, \ldots, K$ **do**

      compute posterior probability Q_{ik}

    end for

  end for

  for each component $k = 1, \ldots, K$ **do**

    compute $\hat{\mu}_k, \hat{\Sigma}_k, \hat{\pi}_k$

  end for

until convergence