

IFT6390

Fondements de l'apprentissage machine

Density estimation

Maximum Likelihood Estimation

Professor: Ioannis Mitliagkas
Slides: Pascal Vincent

Probabilistic approach to learning

- We assumed that the data is generated by an unknown process.
- X, Y is seen as a pair of random variables, distributed according to an unknown probability law $P(X, Y)$.
- X (a vector variable) is itself seen as a set of scalar random variables.

$$P(X, Y) = P(X_{[1]}, \dots, X_{[d]}, Y)$$

Probability

(what you should be familiar with)

- Discrete and continuous random variables
- Joint probability distribution
- Marginal distribution, marginalization
- Conditionnal probability
- Bayes Rule
- Independence

Bayes Rule

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$



Thomas Bayes

1702 - 1761

Homework 1

Operation on distribution

Given a distribution, we may want to

- **Generate data**, i.e. draw samples from this distribution.
- **Compute the probability/likelihood of a configuration** (e.g. knowing the value of some of the variables, after marginalizing the unknown variables).
- **Inference**: *infer* the most likely value or the expectation of some variables given the values of other variables.
- **Learn** the parameters of a distribution **given a data set** (such that the likelihood of the data being generated by this distribution with these parameters is maximized: *maximum likelihood estimation*).

Parametric methods

Objective:

- Estimate the **density** $p(x)$ given an iid dataset $D_n = (X_1, X_2, \dots, X_n)$ drawn from p (p is either a discrete probability or a pdf)

We use a **parametric density** $p(\mathbf{x}) = p(\mathbf{x}|\theta)$

- θ : vector of parameters (ex: probability of landing on heads for a biased coin, mean and variance for a Gaussian distribution, ...)

Maximum likelihood estimation

- To find the *unknown* parameters, we find the value of θ which **maximizes the probability of the data**

Maximum Likelihood Principle

Likelihood of θ with respect to the dataset D_n :

$$p(D_n|\theta) = p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n|\theta) = \prod_{i=1}^n p(\mathbf{x}_i|\theta)$$

Joint distribution over
the \mathbf{x}_i

The \mathbf{x}_i are independent
(i.i.d. assumption)


Log-likelihood of θ with respect to the dataset D_n :

$$\mathcal{L}(\theta) = \log p(D_n|\theta) = \sum_{i=1}^n \log p(\mathbf{x}_i|\theta)$$


Sums are often easier to deal with
than products (e.g., derivatives)

Maximum Likelihood Estimation

- We look for the value of θ (parameters) which maximizes the likelihood:

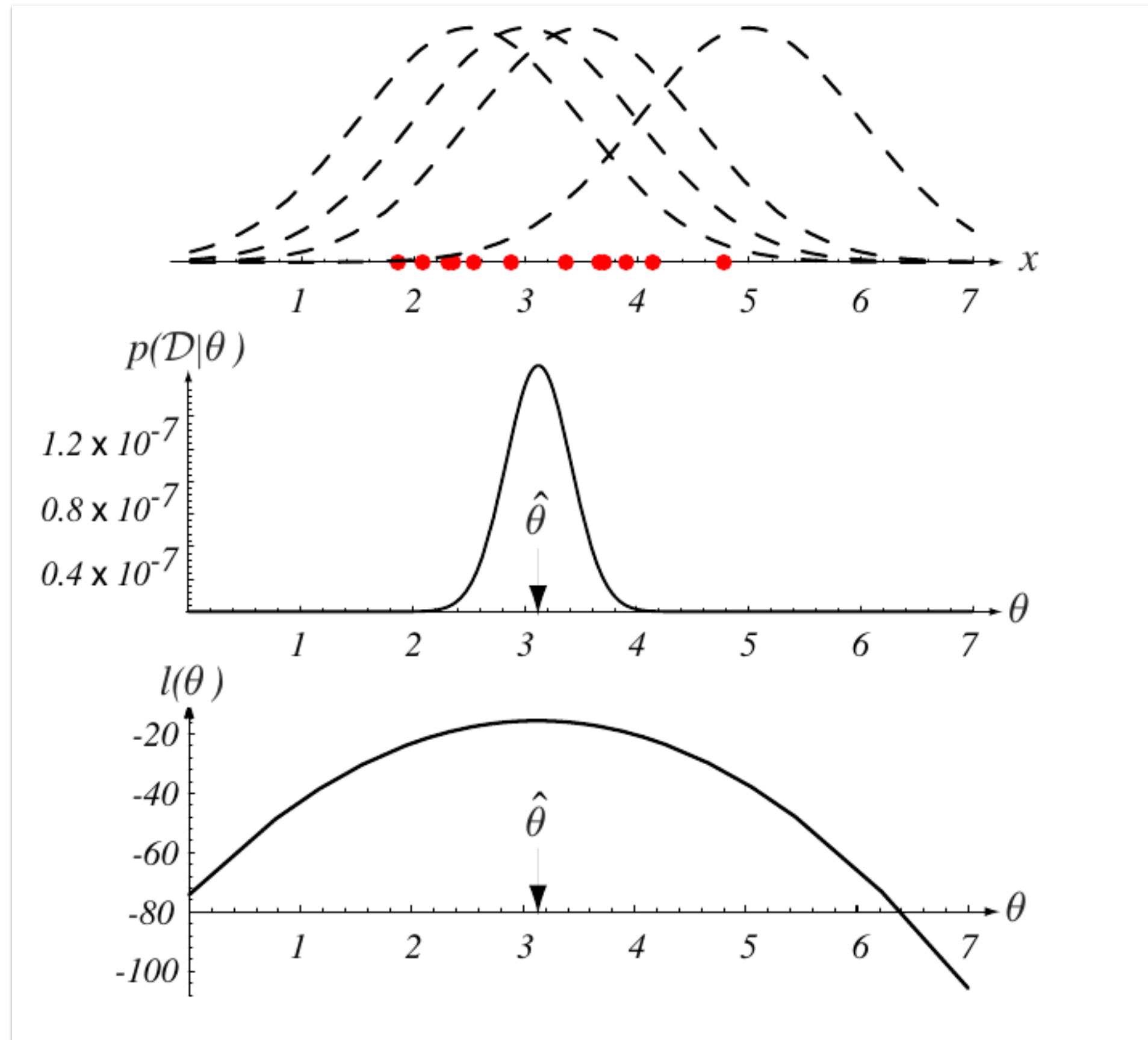
$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} p(D_n|\theta) = \arg \max_{\theta} \prod_{i=1}^n p(\mathbf{x}_i|\theta) \\ &= \arg \max_{\theta} \sum_{i=1}^n \log p(\mathbf{x}_i|\theta) = \arg \max_{\theta} \mathcal{L}(\theta)\end{aligned}$$


Maximum likelihood estimate of θ

- Necessary condition: $\nabla_{\theta} \mathcal{L}(\theta) = \sum_{i=1}^n \nabla_{\theta} \log p(\mathbf{x}_i|\theta) = \mathbf{0}$
- 

(more on this next week)

Maximum Likelihood Estimation



Exercise: Multivariate Gaussian

- We can easily learn the parameters of a Gaussian distribution from a data set:
- The maximum likelihood estimate of μ is the empirical mean (“centroid” of the training points).

$$\mu = \frac{1}{n} \sum_{t=1}^n \mathbf{x}_t$$

- The maximum likelihood estimate of Σ is the empirical covariance matrix:

$$\Sigma_{ij} = \frac{1}{n} \sum_{t=1}^n (\mathbf{x}_{ti} - \mu_i)(\mathbf{x}_{tj} - \mu_j) \quad \text{or} \quad \Sigma = \frac{1}{n} \sum_{t=1}^n (\mathbf{x}_t - \mu)(\mathbf{x}_t - \mu)'$$

=> Try to show these formulas by yourself!