CS 1675 Introduction to Machine Learning
Lecture 8

Density estimation III

Milos Hauskrecht
milos@pitt.edu
5329 Sennott Square

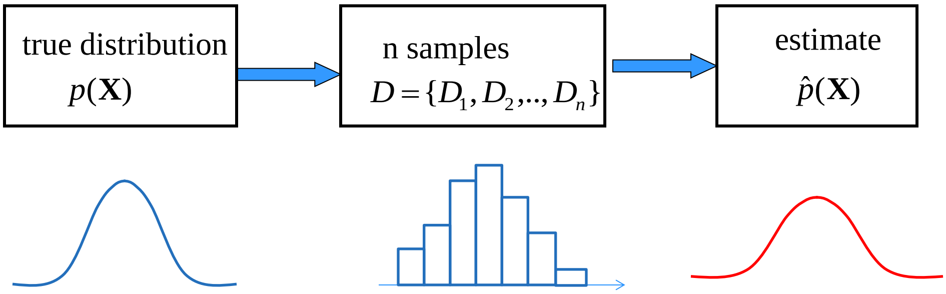
Parametric density estimation

Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$

$D_i = \mathbf{x}_i$ a vector of attribute values

Objective: estimate the model of the underlying probability distribution over variables $\mathbf{X}$, $p(\mathbf{X})$, using examples in D



ML Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta})$ **Data** $D = \{D_1, D_2, .., D_n\}$

- **Maximum likelihood (ML)**
  - Find Θ that maximizes likelihood $p(D \mid \Theta, \xi)$

$\max_{\Theta} \; p(D \mid \Theta, \xi)$

$$P(D \mid \Theta, \xi) = P(D_1, D_2, .., D_n \mid \Theta, \xi)$$
$$= P(D_1 \mid \Theta, \xi) P(D_2 \mid \Theta, \xi) \ldots P(D_n \mid \Theta, \xi)$$
$$= \prod_{i=1}^{n} P(D_i \mid \Theta, \xi)$$

Independent examples

log-likelihood $\log p(D \mid \Theta, \xi) = \sum_{i=1}^{n} \log P(D_i \mid \Theta, \xi)$

$$\Theta_{ML} = \arg\max_{\Theta} \; p(D \mid \Theta, \xi) = \arg\max_{\Theta} \log p(D \mid \Theta, \xi)$$

Bayesian parameter estimation

Bayesian parameter estimation

- Uses the posterior distribution for parameters
- Posterior 'covers' all possible parameter values (and their "weights")

Parameter posterior — Data Likelihood — Parameter prior

$$p(\Theta \mid D, \xi) = \frac{p(D \mid \Theta, \xi)\, p(\Theta \mid \xi)}{p(D \mid \xi)}$$

- **How to use the posterior for modeling** p(X)?

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid D) = \int_{\mathbf{\Theta}} p(X \mid \mathbf{\Theta})\, p(\mathbf{\Theta} \mid D, \xi)\, d\mathbf{\Theta}$$

Posterior Beta distribution

Prior Beta distribution

$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$$

Why to use Beta distribution?

Beta distribution "**fits**" Bernoulli trials, it is called a **conjugate prior**

$$P(D \mid \theta, \xi) = \theta^{N_1}(1-\theta)^{N_2}$$

Posterior distribution is again a Beta distribution !!!!

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) Beta(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

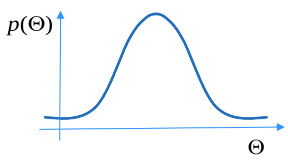
$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_2 + N_2)} \theta^{N_1 + \alpha_1 - 1}(1-\theta)^{N_2 + \alpha_2 - 1}$$

Parameter estimation: MAP

- **Maximum a posteriori probability (MAP)**

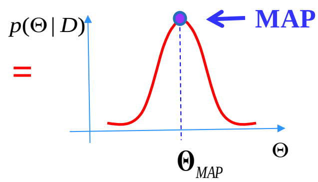
  maximize $p(\mathbf{\Theta} \mid D, \xi)$

Prior on a parameter



$+$ **Data +** $p(D \mid \Theta)$ $=$

Posterior on a parameter



- **MAP**
  - Yields: one set of parameters $\mathbf{\Theta}_{MAP}$ (mode of the posterior)
  - Approximation:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \mathbf{\Theta}_{MAP})$$

Distribution models for random variables

Distribution models covered so far:

- **Bernoulli distribution**
  - **Model for binary random variables**

$$P(x \mid \theta) = \theta^{x}(1-\theta)^{(1-x)}$$

- **Binomial distribution**
  - **Model for order independent sets of binary outcomes**

$$P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1}(1-\theta)^{N-N_1}$$

- **Multinomial distribution**
  - **Model for order independent sets of k-nary outcomes**

$$P(N_1, N_2, \ldots N_k \mid \boldsymbol{\theta}, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \theta_1^{N_1} \theta_2^{N_2} \ldots \theta_k^{N_k}$$

Distribution models for random variables

Models for other types of random variables:

- **Gaussian distribution**
  - **Models of real-valued random variable**
- **Gamma distribution:**
  - **Models of random variables for positive real numbers**
- **Exponential distribution**
  - **Models of random variables for positive real numbers**
- **Poisson distribution**
  - **Models of random variables for nonegative integers**

Conjugate choices of priors for some these distributions:

- **Exponential – Gamma**
- **Poisson – Inverse Gamma**
- **Gaussian - Gaussian (mean) and Wishart (covariance)**

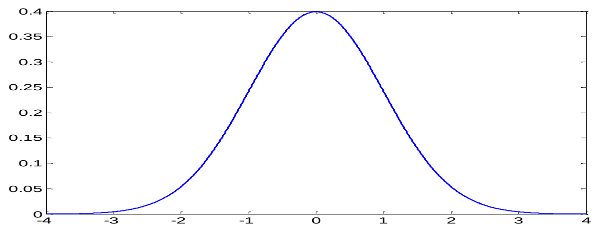
Gaussian (normal) distribution

- **Gaussian:** $x \sim N(\mu, \sigma)$
- **Parameters:** μ - mean

  σ - standard deviation
- **Density function:**

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

- **Example:**



$N(0,1)$

Parameter estimates

- **Loglikelihood**

$$l(D,\mu,\sigma) = \log \prod_{i=1}^{n} p(x_i \mid \mu,\sigma)$$

- **ML estimates of the mean and variance:**

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

  – ML variance estimate is biased

$$E_n(\sigma^2) = E_n\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

- **Unbiased estimate:**

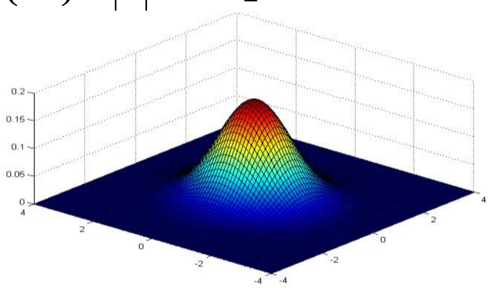
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Multivariate normal distribution

- **Multivariate normal:** $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **Parameters:** $\boldsymbol{\mu}$ - mean
  $\boldsymbol{\Sigma}$ - covariance matrix
- **Density function:**

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

- **Example:**



Partitioned Gaussian Distributions

- **Multivariate Gaussian:**

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- **Example:**

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

Precision matrix

- **What are the distributions for marginals and conditionals?**

$$p(x_a) \qquad p(x_a \mid x_b)$$

Conditionals and Marginals

- **Conditional density:**

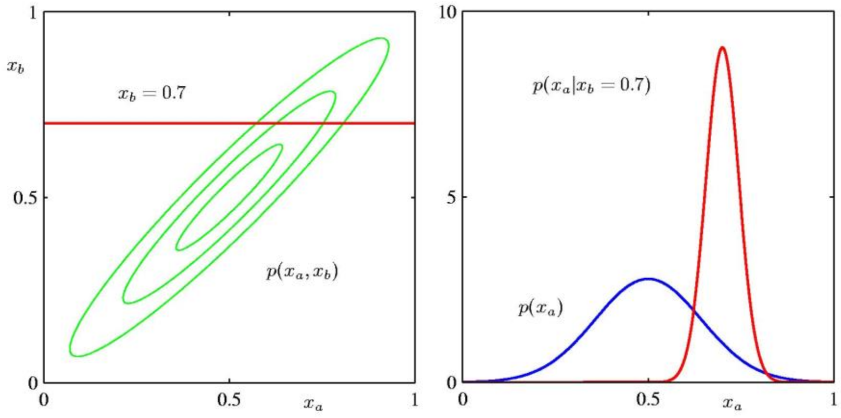
$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\begin{aligned} \boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\ \boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\} \\ &= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\ &= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b) \end{aligned}$$

- **Marginal Density:**

$$\begin{aligned} p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b)\,\mathrm{d}\mathbf{x}_b \\ &= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}) \end{aligned}$$

Conditionals and Marginals



Parameter estimates

- **Loglikelihood** $l(D, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **ML estimates of the mean and covariances:**

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

  – Covariance estimate is biased

$$E_n(\hat{\boldsymbol{\Sigma}}) = E_n\left(\frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T\right) = \frac{n-1}{n}\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}$$

- **Unbiased estimate:**

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

Other distributions

Gamma distribution:

$$p(x \mid a,b) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-\frac{x}{b}} \qquad \text{for} \quad x \in [0,\infty]$$

Exponential distribution:

- A special case of Gamma for a=1

$$p(x \mid b) = \left(\frac{1}{b}\right) e^{-\frac{x}{b}} \qquad \text{for} \quad x \in [0,\infty]$$

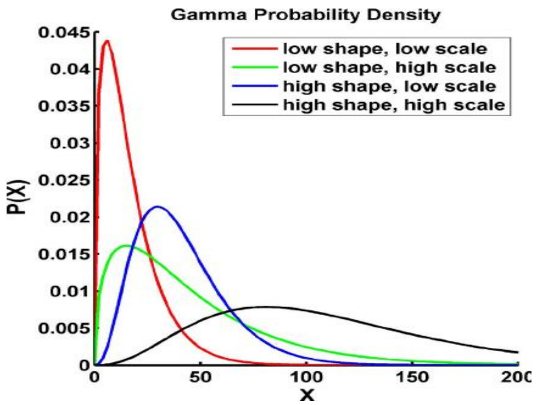
Poisson distribution:

$$p(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{for} \quad x \in \{0,1,2,\ldots\}$$

Gamma distribution

$$p(\lambda \mid a,b) = \frac{1}{\Gamma(a)b^a} \lambda^{a-1} e^{-\frac{\lambda}{b}} \qquad \text{for} \quad \lambda \in [0,\infty]$$

where *a* is the shape and *b* is a scale parameter

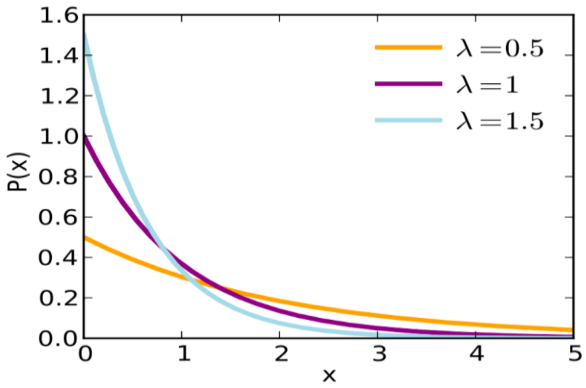


Exponential distribution

$$p(x \mid b) = \left(\frac{1}{b}\right) e^{-\frac{x}{b}} \qquad \text{for} \quad x \in [0, \infty]$$

Alternative parameterization: $p(x \mid \lambda) = \lambda e^{-\lambda x}$

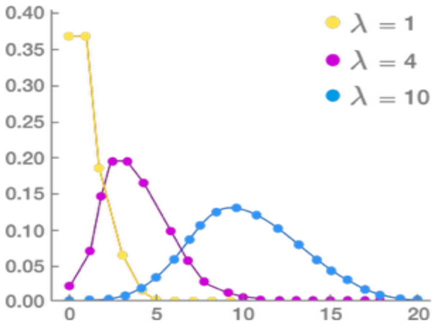
where $\lambda = 1/b$



Poisson distribution

Poisson distribution:

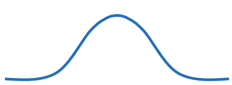
$$p(x \mid \lambda) = \frac{e^{-\lambda} \lambda^{x}}{x!} \qquad \text{for} \quad x \in \{0,1,2,\ldots\}$$



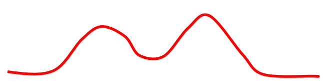
Non-parametric density estimation

Nonparametric Density Estimation

- **Parametric distribution models** are:
  - restricted to specific functional forms, which may not always be suitable;
  - **Example:** modelling a multimodal distribution with a single, unimodal model.



vs



- **Nonparametric approaches:**
  - Do not make any strong assumption about the overall shape of the distribution being modelled.

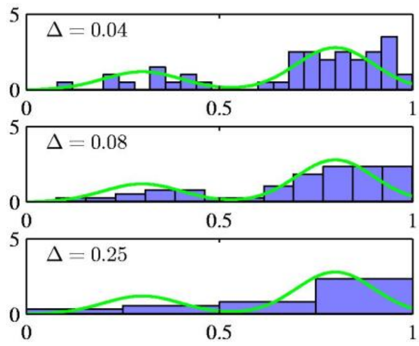
Nonparametric Methods

Histogram methods:

partition the data space into distinct bins with widths Δ_i and count the number of observations, n_i, in each bin.

$$p_i = \frac{n_i}{N\Delta_i}$$

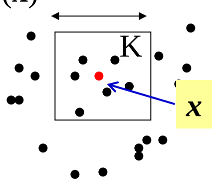
- Often, the same width is used for all bins, $\Delta_i = \Delta$.
- Δ acts as a smoothing parameter.



- Binning does not work well in the in a d-dimensional space,

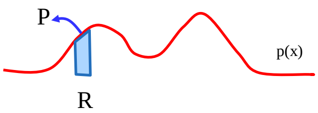
Nonparametric Methods

- Binning does not work well in a d-dimensional space,
  - M bins in each dimension will require M^d bins!
- **Solution:**
  - Build the estimates of p(**x**) by considering the data points in D and how similar (or close) they are to **x**
  - **Example: Parzen window**
    - As if we build a bin dynamically for **x** for which we need to compute p(**x**)

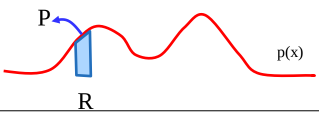


Nonparametric Methods

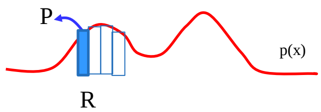
- Assume observations drawn from a density p(x) and consider a small region R containing x such that
$$P = \int_R p(x)dx$$



- The probability that K out of N observations lie inside R is *Bin(K,N,P)* and if N is large
$$K \cong NP$$



If the volume of R, *V*, is sufficiently small, p(x) is approximately constant over R and
$$P \cong p(x)V$$



Thus
$$p(x) = \frac{P}{V}$$

Putting things together we get:

$$p(x) = \frac{K}{NV}$$

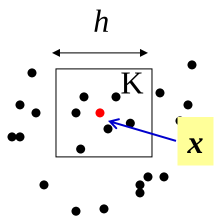
Nonparametric methods: kernel methods

Solution 1: Estimate the probability for **x** based on the fixed volume **V** built around **x**

$$p(x) = \frac{K}{NV}$$

- **Fix V, estimate K from the data**

Example: Parzen window

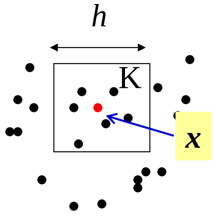


Nonparametric methods: kernel methods

Kernel Density Estimation:

- **Parzen window:** Let R be a hypercube centred on **x** that defines the **kernel function:**

$$k\left(\frac{x-x_n}{h}\right)= \begin{cases} 1 & |(x_i - x_{ni})|/h \le 1/2 \qquad i=1,\ldots D \\ 0 & otherwise \end{cases}$$



•It follows that

$$K=\sum_{n=1}^{N} k\left(\frac{x-x_n}{h}\right)$$

- **and hence**

$$p(x)=\frac{K}{NV}=\frac{1}{Nh^D}\sum_{n=1}^{N} k\left(\frac{x-x_n}{h}\right)$$

Smooth kernels

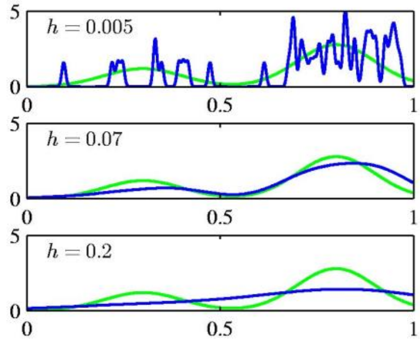
To avoid discontinuities in p(x) because of sharp boundaries we can use a **smooth kernel**, e.g. a Gaussian

$$p(\mathbf{x})=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{\left(2\pi h^2\right)^{D/2}}\exp\left[-\frac{\|\mathbf{x}-\mathbf{x}_n\|}{2h^2}\right]$$

- Any kernel such that

$$k(\mathbf{u}) \ge 0$$
$$\int k(\mathbf{u})d\mathbf{u} = 1$$

- will work.



h acts as a smoother.

Nonparametric Methods: kNN estimation

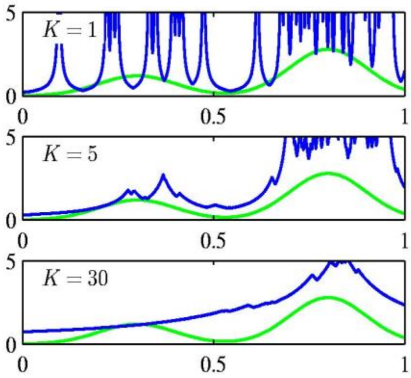
Solution 2: Estimate the probability for **x** based on a fixed count **K** for a variable volume **V** built around **x**

fix K, estimate V from the data

Nearest Neighbour Density Estimation:

Consider a hyper-sphere centred on x and let it grow to a volume, V*, that includes K of the given N data points. Then

$$p(\mathbf{x}) \simeq \frac{K}{NV^{\star}}.$$



K acts as a smoother

Nonparametric vs Parametric Methods

Nonparametric models:

- More flexibility – no density model is needed
- But require storing the entire dataset
- and the computation is performed with all data examples.

Parametric models:

- Once fitted, only parameters need to be stored
- They are much more efficient in terms of computation
- But the model needs to be picked in advance