CS 2750 Machine Learning
Lecture 5

Density estimation

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Density estimations

Topics:

- **Density estimation:** ✓
  - Maximum likelihood (ML)
  - Bayesian parameter estimates
  - MAP
- **Bernoulli distribution.** ✓
- **Binomial distribution** ✓
- **Multinomial distribution** ✓
- **Normal distribution** ✓
- **Exponential family**

Nonparametric family

Parametric density estimation

Parametric density estimation:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$
- **A model of the distribution** over variables in ***X*** with **parameters** Θ : $\hat{p}(\mathbf{X} \mid \Theta)$
- **Data** $D = \{D_1, D_2, .., D_n\}$

Objective:

find parameters Θ such that $p(\mathbf{X} \mid \Theta)$ describes data D the best

Parameter estimation (learning)

- **Maximum likelihood (ML)**

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

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

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

- **Bayesian parameter estimation**
  - **use the posterior density**

$$p(\Theta \mid D, \xi)$$

- **Expected value**

$$\Theta_{EXP} = \int_{\Theta} \Theta p(\Theta \mid D, \xi) d\Theta$$

Exponential family of distribution

Exponential family of distributions

– well behaved distributions with respect to ML and Bayesian updating

Conjugate choices for some of the distributions from the exponential family:

– **Binomial – Beta**
– **Multinomial - Dirichlet**
– **Exponential – Gamma**
– **Poisson – Inverse Gamma**
– **Gaussian - Gaussian (mean) and Wishart (covariance)**

Sequential Bayesian parameter estimation

- **Sequential Bayesian approach**
  – Under the iid the estimates of the posterior can be computed incrementally for a sequence of data points

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

- If we use a conjugate prior we get back the same posterior
- Assume we split the data D in the last element **x** and the rest $p(D \mid \mathbf{\Theta}) = P(x \mid \mathbf{\Theta}) P(D_{n-1} \mid \mathbf{\Theta})$
- **Then:**

A "new" prior

$$p(\Theta \mid D, \xi) = \frac{P(x \mid \mathbf{\Theta}) \overbrace{P(D_{n-1} \mid \mathbf{\Theta}) p(\Theta \mid \xi)}}{\int_{\Theta} P(x \mid \mathbf{\Theta}) P(D_{n-1} \mid \mathbf{\Theta}) p(\Theta \mid \xi) d\Theta}$$

Exponential family

Exponential family:

- all probability mass / density functions that can be written in the exponential normal form

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x})\right]$$

- $\boldsymbol{\eta}$ a vector of **natural (or canonical) parameters**
- $t(\mathbf{x})$ a function referred to as a **sufficient statistic**
- $h(\mathbf{x})$ a function of x (it is less important)
- $Z(\boldsymbol{\eta})$ a normalization constant (a **partition function**)

$$Z(\boldsymbol{\eta}) = \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x})\right\} d\mathbf{x}$$

- Other common form:

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x}) - A(\boldsymbol{\eta})\right] \qquad \log Z(\boldsymbol{\eta}) = A(\boldsymbol{\eta})$$

Exponential family: examples

- **Bernoulli distribution**

$$p(x \mid \pi) = \pi^x (1-\pi)^{1-x}$$

$$= \exp\left\{\log\left(\frac{\pi}{1-\pi}\right) x + \log(1-\pi)\right\}$$

$$= \exp\{\log(1-\pi)\} \exp\left\{\log\left(\frac{\pi}{1-\pi}\right) x\right\}$$

- **Exponential family**

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x})\right]$$

- **Parameters**

$\boldsymbol{\eta} = ?$ $\qquad t(\mathbf{x}) = ?$

$Z(\boldsymbol{\eta}) = ?$ $\qquad h(\mathbf{x}) = ?$

Exponential family: examples

- **Bernoulli distribution**

$$p(x \mid \pi) = \pi^x (1-\pi)^{1-x}$$
$$= \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) x + \log(1-\pi) \right\}$$
$$= \exp\{\log(1-\pi)\} \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) x \right\}$$

- **Exponential family**

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x})\right]$$

- **Parameters**

$$\boldsymbol{\eta} = \log \frac{\pi}{1-\pi} \qquad t(\mathbf{x}) = x$$
$$Z(\boldsymbol{\eta}) = \frac{1}{1-\pi} = 1 + e^{\eta} \qquad h(\mathbf{x}) = 1$$

Exponential family: examples

- **Univariate Gaussian distribution**

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

- **Exponential family**

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(x) \exp\left[\eta^T t(x)\right]$$

- **Parameters**

$$\boldsymbol{\eta} = ? \qquad t(\mathbf{x}) = ?$$
$$Z(\boldsymbol{\eta}) = ? \qquad h(\mathbf{x}) = ?$$

Exponential family: examples

- **Univariate Gaussian distribution**

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

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu}{2\sigma^2} - \log\sigma\right) \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right\}$$

- **Exponential family** $f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(x) \exp\left[\eta^T t(x)\right]$

- **Parameters**

$$\boldsymbol{\eta} = \begin{bmatrix} \mu / 2\sigma^2 \\ -1/2\sigma^2 \end{bmatrix} \qquad t(\mathbf{x}) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

$$Z(\boldsymbol{\eta}) = \exp\left\{\frac{\mu}{2\sigma^2} + \log\sigma\right\} = \exp\left\{-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)\right\}$$

$$h(\mathbf{x}) = 1/\sqrt{2\pi}$$

Exponential family

- **For iid samples, the likelihood of data is**

$$P(D \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} h(\mathbf{x}_i) \exp\left[\boldsymbol{\eta}^T t(\mathbf{x}_i) - A(\boldsymbol{\eta})\right]$$

$$= \left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right] \exp\left[\sum_{i=1}^{n} \boldsymbol{\eta}^T t(\mathbf{x}_i) - A(\boldsymbol{\eta})\right]$$

$$= \left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right] \exp\left[\boldsymbol{\eta}^T \left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right) - nA(\boldsymbol{\eta})\right]$$

- **Important:**
  - the dimensionality of the sufficient statistic remains the same for different sample sizes (that is, different number of examples in D)

Exponential family

- **The log likelihood of data is**

$$l(D,\boldsymbol{\eta}) = \log\left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right]\exp\left[\boldsymbol{\eta}^T\left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right) - nA(\boldsymbol{\eta})\right]$$

$$= \log\left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right] + \left[\boldsymbol{\eta}^T\left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right) - nA(\boldsymbol{\eta})\right]$$

- **Optimizing the loglikelihood**

$$\nabla_{\boldsymbol{\eta}} l(D,\boldsymbol{\eta}) = \left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right) - n\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) \quad = \mathbf{0}$$

- **For the ML estimate it must hold**

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \frac{1}{n}\left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right)$$

Exponential family

- **Rewritting the gradient:**

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \log Z(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \log \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x})\right\} d\mathbf{x}$$

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \frac{\int t(\mathbf{x}) h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x})\right\} d\mathbf{x}}{\int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x})\right\} d\mathbf{x}}$$

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = \int t(\mathbf{x}) h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^T t(\mathbf{x}) - A(\boldsymbol{\eta})\right\} d\mathbf{x}$$

$$\nabla_{\boldsymbol{\eta}} A(\boldsymbol{\eta}) = E(t(\mathbf{x}))$$

- **Result:** $$E(t(\mathbf{x})) = \frac{1}{n}\left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right)$$

- **For the ML estimate, the parameters $\boldsymbol{\eta}$ should be adjusted such that the expectation of the statistic t(x) is equal to the observed sample statistics**

Moments of the distribution

- **For the exponential family**
  - The k-th moment of the statistic corresponds to the k-th derivative of $A(\boldsymbol{\eta})$
  - If x is a component of t(x) then we get the moments of the distribution by differentiating its corresponding natural parameter
- **Example: Bernoulli** $p(x \mid \pi) = \exp\left\{ \log\left(\frac{\pi}{1-\pi}\right) x + \log(1-\pi) \right\}$

  $A(\boldsymbol{\eta}) = \log \frac{1}{1-\pi} = \log(1+e^{\eta})$
- **Derivatives:**

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta} = \frac{\partial}{\partial \eta} \log(1+e^{\eta}) = \frac{e^{\eta}}{(1+e^{\eta})} = \frac{1}{(1+e^{-\eta})} = \pi$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta^2} = \frac{\partial}{\partial \eta} \frac{1}{(1+e^{-\eta})} = \pi(1-\pi)$$

Exponential family of distribution

Bayesian parameter estimate

We have seen conjugate choices for some of the distributions from the exponential family:

- **Binomial – Beta**
- **Multinomial - Dirichlet**
- **Exponential – Gamma**
- **Poisson – Inverse Gamma**
- **Gaussian - Gaussian (mean) and Wishart (covariance)**

Conjugate priors

For any member of the exponential family

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T \mathbf{t}(\mathbf{x})\right]$$

there exists a prior:

$$p(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) = u(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^{\nu} \exp\left[\nu\, \boldsymbol{\eta}^T \boldsymbol{\chi}\right]$$

Such that for n examples, the posterior is

$$p(\boldsymbol{\eta} \mid D, \boldsymbol{\chi}, \nu) \propto g(\boldsymbol{\eta})^{\nu+n} \exp\left[\boldsymbol{\eta}^T \left(\left[\sum_{i=1}^{n} \mathbf{t}(x_i)\right] + \nu \boldsymbol{\chi}\right)\right]$$

Note that:

$$P(D \mid \boldsymbol{\eta}) = \left(\frac{1}{Z(\boldsymbol{\eta})}\right)^n \left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right] \exp\left[\boldsymbol{\eta}^T \left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right)\right]$$

Conjugate priors

For any member of the exponential family

$$f(\mathbf{x} \mid \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})} h(\mathbf{x}) \exp\left[\boldsymbol{\eta}^T \mathbf{t}(\mathbf{x})\right]$$

there exists a prior:

$$p(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) = u(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^{\nu} \exp\left[\nu\, \boldsymbol{\eta}^T \boldsymbol{\chi}\right]$$

Such that for n examples, the posterior is

Pseudo-observation

$$p(\boldsymbol{\eta} \mid D, \boldsymbol{\chi}, \nu) \propto g(\boldsymbol{\eta})^{\nu+n} \exp\left[\boldsymbol{\eta}^T \left(\left[\sum_{i=1}^{n} \mathbf{t}(x_i)\right] + \nu \boldsymbol{\chi}\right)\right]$$

Note that:

$$P(D \mid \boldsymbol{\eta}) = \left(\frac{1}{Z(\boldsymbol{\eta})}\right)^n \left[\prod_{i=1}^{n} h(\mathbf{x}_i)\right] \exp\left[\boldsymbol{\eta}^T \left(\sum_{i=1}^{n} t(\mathbf{x}_i)\right)\right]$$

Nonparametric Methods

- **Parametric distribution models** are:
  - restricted to specific forms, which may not always be suitable;
  - Example: modelling a multimodal distribution with a single, unimodal model.
- **Nonparametric approaches:**
  - make few assumptions about the overall shape of the distribution being modelled.

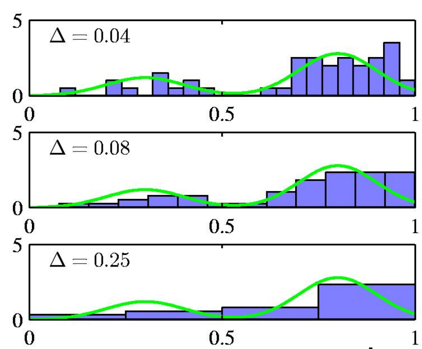
Nonparametric Methods

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.



M^D bins!

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

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

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

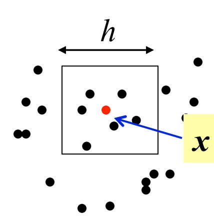
Nonparametric Methods: kernel methods

Kernel Density Estimation:

Fix V, estimate K from the data. Let R be a hypercube centred on **x** and define the kernel function (Parzen window)

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

- **It follows that**
- **and hence** $K = \sum_{n=1}^{N} k\left(\frac{x-x_n}{h}\right)$

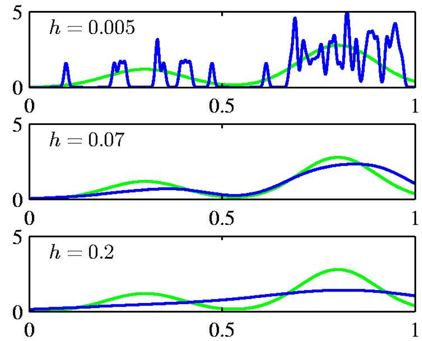


$$p(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{x-x_n}{h}\right)$$

Nonparametric Methods: smooth kernels

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

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



h acts as a smoother.

- Any kernel such that

$$k(\mathbf{u}) \geqslant 0,$$
$$\int k(\mathbf{u})\,\mathrm{d}\mathbf{u} = 1$$

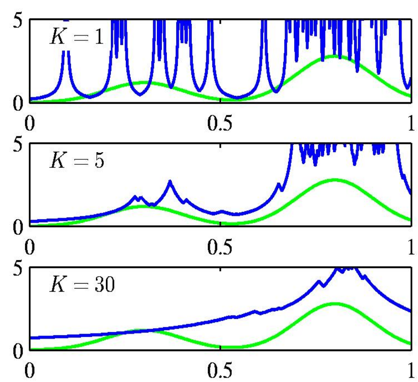
- will work.

Nonparametric Methods: kNN estimation

Nearest Neighbour Density Estimation:

fix K, estimate V from the data. 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