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 \( X = \{X_1, X_2, \ldots, X_n\} \)
- A model of the distribution over variables in \( X \) with parameters \( \Theta : \hat{p}(X | \Theta) \)
- Data \( D = \{D_1, D_2, \ldots, D_n\} \)

**Objective:**
find parameters \( \Theta \) such that \( p(X | \Theta) \) describes data \( D \) the best

Parameter estimation (learning)

- **Maximum likelihood (ML)**
  \[ \Theta_{ML} = \arg \max_{\Theta} p(D | \Theta, \xi) \]

- **Maximum a posteriori probability (MAP)**
  \[ \Theta_{MAP} = \arg \max_{\Theta} p(\Theta | D, \xi) \]

- **Bayesian parameter estimation**
  - use the posterior density
  \[ p(\Theta | D, \xi) \]
  - **Expected value**
  \[ \Theta_{EXP} = \int_{\Theta} \Theta p(\Theta | D, \xi) d\Theta \]
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

\[
p(\Theta | D, \tilde{\xi}) = \frac{p(D | \Theta, \tilde{\xi}) p(\Theta | \tilde{\xi})}{\int p(D | \Theta, \tilde{\xi}) p(\Theta | \tilde{\xi}) d\Theta}
\]

– 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
\[
p(D | \Theta) = P(x | \Theta) P(D_{n-1} | \Theta)
\]

**Then:**

\[
p(\Theta | D, \tilde{\xi}) = \frac{P(x | \Theta) P(D_{n-1} | \Theta) p(\Theta | \tilde{\xi})}{\int P(x | \Theta) P(D_{n-1} | \Theta) p(\Theta | \tilde{\xi}) d\Theta}
\]

A “new” prior
Exponential family

Exponential family:
- all probability mass / density functions that can be written in the exponential normal form
  \[ f(x \mid \eta) = \frac{1}{Z(\eta)} h(x) \exp[\eta^T t(x)] \]
- \( \eta \) a vector of natural (or canonical) parameters
- \( t(x) \) a function referred to as a sufficient statistic
- \( h(x) \) a function of \( x \) (it is less important)
- \( Z(\eta) \) a normalization constant (a partition function)
  \[ Z(\eta) = \int h(x) \exp\{\eta^T t(x)\} dx \]
- Other common form:
  \[ f(x \mid \eta) = h(x) \exp[\eta^T t(x) - A(\eta)] \]
  \[ \log Z(\eta) = A(\eta) \]

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
  \[ f(x \mid \eta) = \frac{1}{Z(\eta)} h(x) \exp[\eta^T t(x)] \]
- Parameters
  \[ \eta = ? \quad t(x) = ? \]
  \[ Z(\eta) = ? \quad h(x) = ? \]
**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**
  \[ f(x \mid \eta) = \frac{1}{Z(\eta)} h(x) \exp[\eta^T t(x)] \]

- **Parameters**
  \[ \eta = \log \frac{\pi}{1 - \pi} \]
  \[ t(x) = x \]
  \[ Z(\eta) = \frac{1}{1 - \pi} = 1 + e^{\eta} \]
  \[ h(x) = 1 \]

---

**Exponential family: examples**

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

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

- **Parameters**
  \[ \eta = ? \]
  \[ t(x) = ? \]
  \[ Z(\eta) = ? \]
  \[ h(x) = ? \]
Exponential family: examples

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

- Exponential family
  \[ f(x \mid \eta) = \frac{1}{Z(\eta)} h(x) \exp[\eta^T t(x)] \]

- Parameters
  \[ \eta = \begin{bmatrix} \mu / 2\sigma^2 \\ -1 / 2\sigma^2 \end{bmatrix} \]
  \[ t(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \]
  \[ Z(\eta) = \exp\left\{ \frac{\mu}{2\sigma^2} + \log \sigma \right\} = \exp\left\{ -\frac{\eta^2}{4\eta^2} - \frac{1}{2} \log(-2\eta) \right\} \]
  \[ h(x) = 1 / \sqrt{2\pi} \]

---

Exponential family

- For iid samples, the likelihood of data is
  \[ P(D \mid \eta) = \prod_{i=1}^n p(x_i \mid \eta) = \prod_{i=1}^n h(x_i) \exp[\eta^T t(x_i) - A(\eta)] \]
  \[ = \left[ \prod_{i=1}^n h(x_i) \right] \exp \left[ \sum_{i=1}^n \eta^T t(x_i) - A(\eta) \right] \]
  \[ = \left[ \prod_{i=1}^n h(x_i) \right] \exp \left[ \eta^T \left( \sum_{i=1}^n t(x_i) \right) - nA(\eta) \right] \]

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

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

• Optimizing the loglikelihood
  \[ \nabla_\eta l(D, \eta) = \left( \sum_{i=1}^{n} t(x_i) \right) - n\nabla_\eta A(\eta) = 0 \]

• For the ML estimate it must hold
  \[ \nabla_\eta A(\eta) = \frac{1}{n} \left( \sum_{i=1}^{n} t(x_i) \right) \]

---

Exponential family

• Rewriting the gradient:
  \[ \nabla_\eta A(\eta) = \nabla_\eta \log Z(\eta) = \nabla_\eta \log \int h(x) \exp \left\{ \eta^T t(x) \right\} dx \]

  \[ \nabla_\eta A(\eta) = \int t(x) h(x) \exp \left\{ \eta^T t(x) \right\} dx \]

  \[ \nabla_\eta A(\eta) = \frac{\int t(x) h(x) \exp \left\{ \eta^T t(x) \right\} dx}{\int h(x) \exp \left\{ \eta^T t(x) \right\} dx} \]

  \[ = \int t(x) h(x) \exp \left\{ \eta^T t(x) - A(\eta) \right\} dx \]

  \[ \nabla_\eta A(\eta) = E(t(x)) \]

• Result:
  \[ E(t(x)) = \frac{1}{n} \left( \sum_{i=1}^{n} t(x_i) \right) \]

• For the ML estimate, the parameters \( \eta \) 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 $A(\eta)$
  - If $x$ is a component of $t(x)$ then we get the moments of the distribution by differentiating its corresponding natural parameter

- Example: Bernoulli $p(x | \pi) = \exp \left\{ \log \left( \frac{\pi}{1-\pi} \right) x + \log(1 - \pi) \right\}$
  
  $A(\eta) = \log \frac{1}{1 - \pi} = \log(1 + e^\eta)$

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

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

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

there exists a prior:

\[ p(\eta \mid \chi, \nu) = u(\chi, \nu) g(\eta)^\nu \exp[\nu^T \chi] \]

Such that for n examples, the posterior is

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

Note that:

\[ P(D \mid \eta) = \left( \frac{1}{Z(\eta)} \right)^n \prod_{i=1}^n \left[ h(x_i) \right] \exp\left[ \eta^T \left( \sum_{i=1}^n t(x_i) \right) \right] \]
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.

Histogram methods:
partition the data space into distinct bins with widths \( \Delta_i \) 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, \( \Delta_i = \Delta \).
- \( \Delta \) acts as a smoothing parameter.
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 \( \text{Bin}(K,N,P) \) and if \( N \) is large

\[
K \approx NP
\]

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

\[
P = p(x)V
\]

Thus

\[
p(x) = \frac{P}{V}
\]

\[
p(x) = \frac{K}{NV}
\]

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)

\[
k \left( \frac{x - x_n}{h} \right) = \begin{cases} 1 & |(x_i - x_n)| / h \leq 1/2 \quad i = 1, \ldots, D \\ 0 & \text{otherwise} \end{cases}
\]

• 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

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

- Any kernel such that
  \[
h(u) \geq 0, \quad \int h(u) \, du = 1
\]
- \( h \) acts as a smoother.

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

\[
p(x) \simeq \frac{K}{NV^*}.
\]