CS 3750 Machine Learning

Lecture 2

Density estimation

Milos Hauskrecht
milos@cs.pitt.edu
Sennott Square, x4-8845

http://www.cs.pitt.edu/~milos/courses/cs3750/

Density estimation

Data: $D = \{D_1, D_2, .., D_n\}$
$D_i = \mathbf{x}_i$ a vector of attribute values

Attributes:

- modeled by random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ with:
  - **Continuous values**
  - **Discrete values**

E.g. ***blood pressure*** with numerical values
or ***chest pain*** with discrete values
[no-pain, mild, moderate, strong]

Underlying true probability distribution:

$p(\mathbf{X})$

Density estimation

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

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

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

true distribution $p(\mathbf{X})$ → n samples $D = \{D_1, D_2, .., D_n\}$ → estimate $\hat{p}(\mathbf{X})$

Standard (iid) assumptions: Samples
- **are independent of each other**
- **come from the same (identical) distribution (fixed $p(\mathbf{X})$)**

Density estimation

Types of density estimation:

Parametric
- the distribution is modeled using a set of parameters Θ

$$p(\mathbf{X} \mid \Theta)$$

- **Example:** mean and covariances of multivariate normal
- **Estimation:** find parameters $\hat{\Theta}$ that fit the data D the best

Non-parametric
- The model of the distribution utilizes all examples in D
- As if all examples were parameters of the distrtribution
- **Examples:** Nearest-neighbor

Semi-parametric

Parametric density estimation

Parametric density estimation

Basic settings:

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

Objective: find parameters $\hat{\Theta}$ that describe $p(\mathbf{X} \mid \Theta)$ the best

Parameter learning

What is the best set of parameters?

- **Maximum likelihood (ML) estimates**

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

  ξ - represents prior (background) knowledge

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

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

  Selects the mode of the posterior

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

Parameter learning

- **Both ML or MAP pick one parameter value**
  - Is it always the best solution?
- **Bayesian approach**
  - Remedies the limitation of one choice
  - Keeps and uses complete posterior distribution $p(\Theta \mid D, \xi)$
  - Optimization is replaced with integration
- **How is it used? Assume we want:** $P(\mathbf{x} \mid D, \xi)$
  - Consider all parameter settings and averages the result

$$P(\mathbf{x} \mid D, \xi) = \int_\theta P(\mathbf{x} \mid \theta, \xi) p(\theta \mid D, \xi) d\theta$$

  - **Example:** predict the result of the outcome x=1

$$P(x=1 \mid D, \xi)$$

Bernoulli distribution.

Outcomes: x_i with values 0 or 1 (head or tail)
Data: D a sequence of outcomes x_i

Model: probability of an outcome 1 θ
probability of 0 $(1-\theta)$

$$P(x_i \mid \theta) = \theta^{x_i} (1-\theta)^{(1-x_i)}$$ **Bernoulli distribution**

Maximum likelihood (ML) estimate.

Likelihood of data:

$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)}$$

Maximum likelihood estimate

$$\theta_{ML} = \arg\max_{\theta} P(D \mid \theta, \xi)$$

Optimize log-likelihood

$$l(D,\theta) = \log P(D \mid \theta, \xi) = \log \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} =$$

$$\sum_{i=1}^{n} x_i \log\theta + (1-x_i)\log(1-\theta) = \log\theta \underbrace{\sum_{i=1}^{n} x_i}_{N_1} + \log(1-\theta) \underbrace{\sum_{i=1}^{n}(1-x_i)}_{N_2}$$

N_1 - number of 1s seen $\quad$ N_2 - number of 0s seen

Maximum likelihood (ML) estimate.

Optimize log-likelihood

$$l(D,\theta) = N_1 \log\theta + N_2 \log(1-\theta)$$

Set derivative to zero

$$\frac{\partial l(D,\theta)}{\partial \theta} = \frac{N_1}{\theta} - \frac{N_2}{(1-\theta)} = 0$$

Solving $\quad \theta = \dfrac{N_1}{N_1 + N_2}$

ML Solution: $\quad \theta_{ML} = \dfrac{N_1}{N} = \dfrac{N_1}{N_1 + N_2}$

Maximum a posteriori estimate

Maximum a posteriori estimate

– Selects the mode of the posterior distribution

$$\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D, \xi)$$

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) p(\theta \mid \xi)}{P(D \mid \xi)} \quad \textbf{(via Bayes rule)}$$

$P(D \mid \theta, \xi)$ - is the likelihood of data

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

$p(\theta \mid \xi)$ - is the prior probability on θ

How to choose the prior probability?

Prior distribution

Choice of 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?

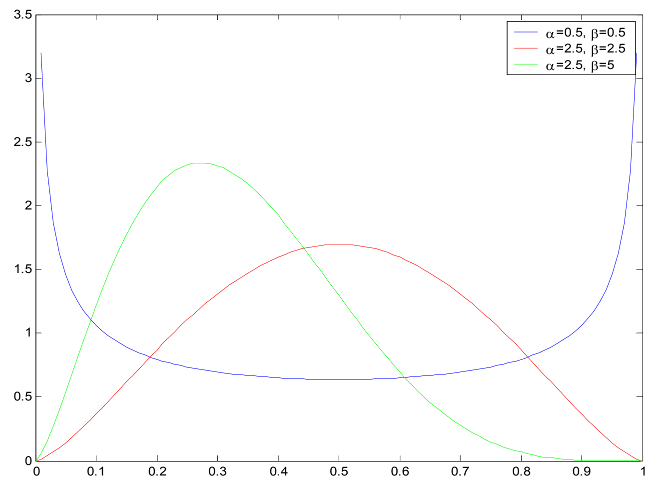
Beta distribution "**fits**" binomial sampling - **conjugate choices**

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

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

MAP Solution: $$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$

Beta distribution



Bayesian approach

- **Posterior probability:**

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

- **Probability of an outcome** $x = 1$ **in the next trial**

$$P(x=1 \mid D, \xi) = \int_0^1 P(x=1 \mid \theta, \xi) p(\theta \mid D, \xi) d\theta$$

$$= \int_0^1 \theta p(\theta \mid D, \xi) d\theta = E(\theta)$$

- **Equivalent to the expected value of the parameter**
  - expectation is taken with regard to the posterior distribution

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

Bayesian learning

Expected value of the parameter

$$E(\theta) = \int_0^1 \theta Beta(\theta \mid \eta_1, \eta_2) d\theta = \int_0^1 \theta \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \theta^{\eta_1 - 1} (1-\theta)^{\eta_2 - 1} d\theta$$

$$= \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \int_0^1 \theta^{\eta_1} (1-\theta)^{\eta_2 - 1} d\theta$$

$$= \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \frac{\Gamma(\eta_1 + 1)\Gamma(\eta_2)}{\Gamma(\eta_1 + \eta_2 + 1)} \underbrace{\int_0^1 Beta(\eta_1 + 1, \eta_2) d\theta}_{1}$$

$$= \frac{\eta_1}{\eta_1 + \eta_2}$$

Note: $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$

Expected value

- **Predictive probability of an outcome $x = 1$ in the next trial**

  $P(x = 1 \mid D, \xi) = E(\theta)$

- **Substituting the results for**

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

- **We get**

  $$P(x = 1 \mid D, \xi) = E(\theta) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$$

- **Instead of MAP and ML choice of the parameter we can use the expected value of the parameter**

  $\hat{\theta} = E(\theta)$

Binomial distribution.

Data: D a set of order-independent outcomes with two possible values – 0 or 1 (head or tail)

N_1 - number of heads seen N_2 - number of tails seen

We treat *D* as a multi-set !!!

Model: probability of a 1 θ

probability of a 0 $(1-\theta)$

Probability of an outcome

$$P(D \mid \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_2}$$ **Binomial distribution**

Maximum likelihood (ML) estimate.

Likelihood of data:

$$P(D \mid \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_2} = \frac{N!}{N_1! N_2!} \theta^{N_1} (1-\theta)^{N_2}$$

Log-likelihood

$$l(D,\theta) = \log \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_2} = \underline{\log \frac{N!}{N_1! N_2!}} + N_1 \log\theta + N_2 \log(1-\theta)$$

Constant from the point of optimization !!!

ML Solution: $\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$

The same as for Bernoulli and D **with iid** sequence of examples

Maximum a posteriori estimate

MAP estimate $\theta_{MAP} = \arg\max_{\theta} p(\theta \mid D, \xi)$

$$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(via Bayes rule)**

Prior choice

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

Likelihood

$$P(D \mid \theta) = \frac{\Gamma(N_1 + N_2)}{\Gamma(N_1)\Gamma(N_2)} \theta^{N_1}(1-\theta)^{N_2}$$

Posterior

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

MAP estimate

$$\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}$$

Bayesian learning, expectation

The result is the same as for Bernoulli distribution

$$E(\theta) = \int_0^1 \theta Beta(\theta \mid \eta_1, \eta_2) d\theta = \frac{\eta_1}{\eta_1 + \eta_2}$$

Expected value of the parameter

$$E(\theta) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$$

Predictive probability of an event **x**=1

$$E(\theta) = P(x = 1 \mid \theta, \xi) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$$

Multinomial distribution

A kind of multi-way coin toss (roll of dice)

- **Data:** a set of N outcomes (multi-set)

  N_i - a number of times an outcome i has been seen

Model parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots \theta_k)$ **s.t.** $\sum_{i=1}^{k} \theta_i = 1$

θ_i - probability of an outcome i

Probability of data (likelihood)

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

Multinomial distribution

ML estimate:

$$\theta_{i,ML} = \frac{N_i}{N}$$

MAP estimate

Choice of prior: Dirichlet distribution

$$Dir(\boldsymbol{\theta} \mid \alpha_1, .., \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1}$$

Dirichlet is the conjugate choice for multinomial

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

Posterior distribution

$$p(\boldsymbol{\theta} \mid D, \xi) = \frac{P(D \mid \boldsymbol{\theta}, \xi) Dir(\boldsymbol{\theta} \mid \alpha_1, \alpha_2, ..\alpha_k)}{P(D \mid \xi)} = Dir(\boldsymbol{\theta} \mid \alpha_1 + N_1, .., \alpha_k + N_k)$$

MAP estimate:

$$\theta_{i,MAP} = \frac{\alpha_i + N_i - 1}{\sum_{i=1,..k} (\alpha_i + N_i) - k}$$

Bayesian learning

The result is analogous to the result for binomial

$$E(\boldsymbol{\theta}) = \oint_{0 \le \theta_i \le 1, \sum \theta_i = 1} \boldsymbol{\theta} \; Dir(\boldsymbol{\theta} \mid \boldsymbol{\eta}) d\boldsymbol{\theta} = \left(\frac{\eta_1}{\eta_1 + \eta_2 + \eta_k}, \ldots \frac{\eta_i}{\eta_1 + \eta_2 + \eta_k}, \ldots \frac{\eta_k}{\eta_1 + \eta_2 + \eta_k} \right)$$

Bayesian estimate substitutes posterior

$$E(\boldsymbol{\theta}) = \left(\frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \ldots + \alpha_k + N_k} \ldots \frac{\alpha_i + N_i}{\alpha_1 + N_1 + \ldots + \alpha_k + N_k} \ldots \frac{\alpha_k + N_k}{\alpha_1 + N_1 + \ldots + \alpha_k + N_k} \right)$$

Represents the predictive probability of an event x=i

$$P(x = i \mid \boldsymbol{\theta}, \xi) = \frac{\alpha_i + N_i}{\alpha_1 + N_1 + \ldots + \alpha_k + N_k}$$

Other distributions

The same ideas can be applied to other distributions

– Typically we choose distributions that behave in a nice way so that the computations lead to "nice" solutions

- **Exponential family of distributions**

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

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

Distributions from the exponential family

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

$$p(x \mid b) = \left(\frac{1}{b}\right) e^{-\frac{x}{b}}$$

Poisson distribution:

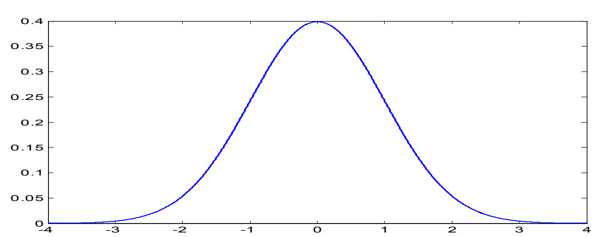
$$p(x \mid \lambda) = \frac{e^{-\lambda} \lambda^{x}}{x!} \qquad \text{for} \quad x \in \{0,1,2,\ldots\}$$

Gaussian (normal) distribution

- **Gaussian:** $x \sim N(\mu, \sigma)$
- **Parameters:** μ - mean
  σ - standard deviation
- **Density function:**

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

- **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:**

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