CS 3710 Advanced Topics in AI
Lecture 17

Density estimation

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

Administration

Midterm:

- **A take-home exam (1 week)**
- **Due on Wednesday, November 2, 2005 before the class**
- **Depends on the material covered so far:**
  - Exact inferences
  - Monte-Carlo sampling
  - Variational approximation
- **You will be evaluated on the correctness and clarity of your answers**
  - Be neat and explain clearly your notations and solutions

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:

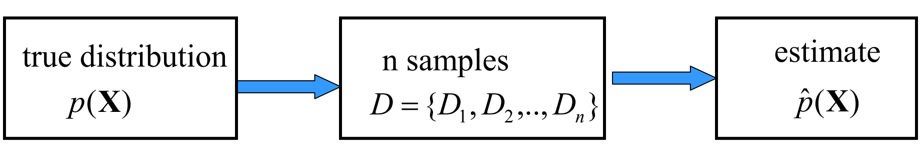
$p(\mathbf{X})$

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



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 distribution
- **Examples:** Nearest-neighbor

Semi-parametric

Parametric density estimation

P**arametric 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

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

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

N_1 - number of 1s seen 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 $\theta = \dfrac{N_1}{N_1 + N_2}$

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

Maximum a posteriori estimate

Maximum a posteriori estimate

- Selects the mode of the posterior distribution

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

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

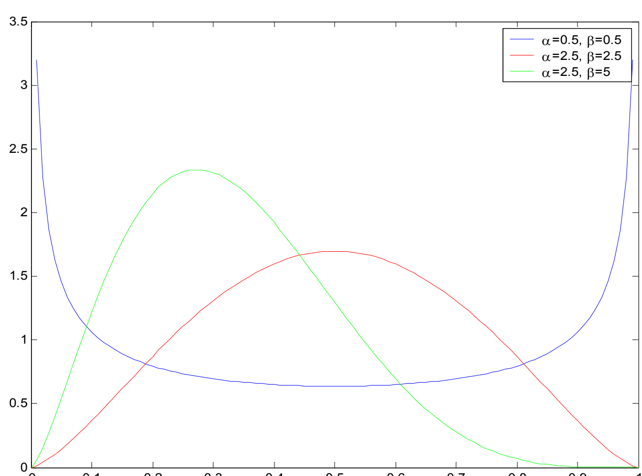
Beta distribution "**fits**" binomial sampling - **conjugate choices**

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

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

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**

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