CS 2750 Machine Learning
Lecture 4

Density estimation

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

Announcements

Homework 1

- Due on Wednesday before the class
- **Reports:** hand in before the class
- **Programs:** submit electronically

Collaborations on homeworks:

- You may discuss material with your fellow students, but the report and programs should be written individuall

Outline

Outline:
- Density estimation.
- Bernoulli distribution.
- Binomial

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 Θ describing data D

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

Learning via parameter estimation

In this lecture we consider **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 :

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

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

Objective: find the description of parameters Θ so they fit the observed data

Parameter estimation.

- **Maximum likelihood (ML)**

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

  – yields: one set of parameters Θ_{ML}

  – the target distribution is approximated as:

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

- **Bayesian parameter estimation**

  – uses the posterior distribution over possible parameters

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

  – Yields: all possible settings of Θ (and their “weights”)

  – The target distribution is approximated as:

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

Parameter estimation.

Other possible criteria:

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

  maximize $p(\mathbf{\Theta} \mid D, \xi)$ (mode of the posterior)

  – Yields: one set of parameters $\mathbf{\Theta}_{MAP}$

  – Approximation:

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

- **Expected value of the parameter**

  $\hat{\mathbf{\Theta}} = E(\mathbf{\Theta})$ (mean of the posterior)

  – Expectation taken with regard to posterior $p(\mathbf{\Theta} \mid D, \xi)$

  – Yields: one set of parameters

  – Approximation:

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

Example: Bernoulli distribution.

Coin example: we have a coin that can be biased

Outcomes: two possible values -- head or tail

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Objective:

We would like to estimate the probability of a **head** $\hat{\theta}$

Probability of an outcome x_i

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

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

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

N_1 - number of heads seen N_2 - number of tails 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: $\theta_{ML} = \dfrac{N_1}{N} = \dfrac{N_1}{N_1 + N_2}$

Bayesian parameter estimate

Posterior distribution

$$p(\theta \mid D,\xi) = \frac{P(D \mid \theta,\xi)\, p(\theta \mid \xi)}{P(D \mid \xi)}$$ **(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 the prior: Beta distribution

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

$$\Gamma(a) = \int_{-\infty}^{\infty} x^{a-1}e^{-x}dx$$ - Gamma function

Why Beta?

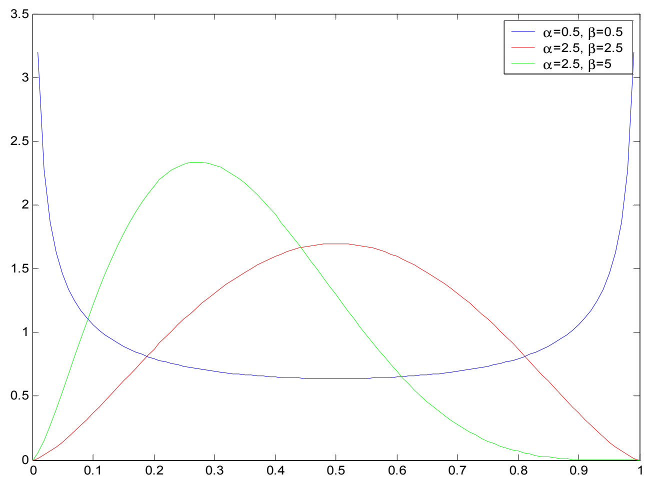
Beta distribution "**fits**" binomial sampling - **conjugate choices**

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

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

Parameters: $\alpha_1+N_1,\alpha_2+N_2$

Beta distribution



MAP solution

Maximum a posteriori estimate

– Selects the mode of **the posterior distribution**

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

- **MAP solution for Beta prior**

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

MAP Solution:

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

Bayesian framework

- **Both ML or MAP estimates pick one value of the parameter**
  - **Assume:** there are two different parameter settings that are close in terms of their probability values. Using only one of them may introduce a strong bias, if we use them, for example, for predictions.
- **Bayesian parameter estimate**
  - Remedies the limitation of one choice
  - Uses all possible parameter values
  - Where $p(\theta \mid D, \xi) \approx Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$
- **The posterior can be used to define** $\hat{p}(\mathbf{X})$ **:**

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

Bayesian framework

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

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

Posterior density

$$P(x = 1 \mid D, \xi) = \int_0^1 P(x = 1 \mid \theta, \xi) \overbrace{p(\theta \mid D, \xi)}^{\text{Posterior density}} d\theta$$

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

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

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

Expected value of the parameter

How to obtain the expected value?

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

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

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

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

Note: $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for integer values of α

Expected value of the parameter

- **Substituting the results for the posterior:**

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

- **We get** $$E(\theta) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}$$

- **Note that the mean of the posterior is yet** another "reasonable" parameter choice:

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

Binomial distribution.

Example problem: a biased coin

Outcomes: two possible values -- head or tail

Data: D a set of order-independent outcomes

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

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

Model: probability of a head θ

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

Objective:

We would like to estimate the probability of a **head** $\hat{\theta}$

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

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

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

Posterior density

Posterior density

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

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

Expected value of the parameter

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 event x=1

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