Density estimation

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

Parametric density estimation

Parametric density estimation:
• A set of random variables \( X = \{ X_1, X_2, \ldots, X_d \} \)
• 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**
  \[ \Theta_{EXP} = \int_{\Theta} \Theta p(\Theta | D, \xi) d\Theta \]

---

Probability of a binary outcome

**Data:** \( D \) a sequence of outcomes \( x_i \) such that
- **head** \( x_i = 1 \)
- **tail** \( x_i = 0 \)

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

**Assume:** we know the probability \( \theta \)
**Probability of an outcome of a coin flip** \( x_i \)

\[ P(x_i | \theta) = \theta^{x_i} (1 - \theta)^{1-x_i} \]

- Combines the probability of a head and a tail
- So that \( x_i \) is going to pick its correct probability
- Gives \( \theta \) for \( x_i = 1 \)
- Gives \( (1 - \theta) \) for \( x_i = 0 \)
The goodness of fit to the data

Learning: we do not know the value of the parameter $\theta$

Our learning goal:
- Find the parameter $\theta$ that fits the data $D$ the best?

One solution to the “best”: Maximize the likelihood

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

Intuition:
- more likely are the data given the model, the better is the fit

Assuming a sequence of $n$ Bernoulli trials:

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

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 (the same as maximizing likelihood)

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

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

$N_1$ - number of heads seen \quad $N_2$ - number of tails seen
Maximum a posteriori estimate

- Selects the mode of the posterior distribution

\[ \theta_{MAP} = \arg \max_\theta p(\theta | D, \xi) \]

Likelihood of data

\[ p(\theta | D, \xi) = \frac{P(D | \theta, \xi) p(\theta | \xi)}{P(D | \xi)} \]

(via Bayes rule)

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

\[ p(\theta | \xi) \] - is the prior probability on \( \theta \)

How to choose the prior probability?

Prior distribution

Choice of prior: Beta distribution

\[ p(\theta | \xi) = \text{Beta}(\theta | \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} \]

\( \Gamma(x) \) - a Gamma function \( \Gamma(x) = (x-1)! \)

For integer values of \( x \) \( \Gamma(n) = (n-1)! \)

Why to use Beta distribution?

Beta distribution "fits" Bernoulli trials - conjugate choices

\[ P(D | \theta, \xi) = \theta^{N_1}(1-\theta)^{N_2} \]

Posterior distribution is again a Beta distribution

\[ p(\theta | D, \xi) = \frac{P(D | \theta, \xi) \text{Beta}(\theta | \alpha_1, \alpha_2)}{P(D | \xi)} = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2) \]
Posterior distribution

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

Maximum a posterior probability

**Maximum a posteriori estimate**

- Selects the mode of the posterior distribution

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

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

\[ = \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_2 + N_2)} \theta^{\alpha_1 + N_1 - 1}(1-\theta)^{\alpha_2 + N_2 - 1} \]

**Notice** that parameters of the prior act like counts of heads and tails

(sometimes they are also referred to as **prior counts**)

**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
– Keeps all possible parameter values
– Where \( p(\theta \mid D, \xi) \approx \text{Beta}(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2) \)

• The posterior can be used to define \( p(A \mid D) \):

\[
p(A \mid D) = \int p(A \mid \Theta) p(\Theta \mid D, \xi) d\Theta
\]

Binomial distribution

Example problem: a biased coin

Outcomes: two possible values -- head or tail

Data: a set of order-independent outcomes for \( N \) trials

\( N_1 \) - number of heads seen \( \quad \) \( N_2 \) - number of tails seen can be calculated from the trial data !!!

Model: probability of a head \( \theta \)

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

Probability of an outcome

\[
P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1} (1 - \theta)^{N - N_1} \quad \text{Binomial distribution}
\]

Objective:

We would like to estimate the probability of a head \( \hat{\theta} \)
Binomial distribution

Binomial distribution:

Maximum likelihood (ML) estimate.

Likelihood of data:
\[ P(D | \theta) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_2} = \frac{N!}{N_1! \cdot N_2!} \theta^{N_1} (1-\theta)^{N_2} \]

Log-likelihood
\[ l(D, \theta) = \log \left( \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_2} \right) = \log \frac{N!}{N_1! \cdot N_2!} + N_1 \log \theta + N_2 \log(1-\theta) \]

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

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
\[ p(\theta \mid D, \xi) = \text{Beta}(\alpha_1 + N_1, \alpha_2 + N_2) \]

MAP estimate
\[ \theta_{\text{MAP}} = \arg \max_\theta p(\theta \mid D, \xi) \]
\[ \theta_{\text{MAP}} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2} \]

Expected value of the parameter

The result is the same as for Bernoulli distribution
\[ E(\theta) = \int_0^1 \theta \text{Beta}(\theta \mid \eta_1, \eta_2) d\theta = \frac{\eta_1}{\eta_1 + \eta_2} \]

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

Example: 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: $\theta = (\theta_1, \theta_2, \ldots \theta_k)$ s.t. $\sum_{i=1}^{k} \theta_i = 1$

$\theta_i$ - probability of an outcome i

Probability of data (likelihood)

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

ML estimate:

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

Posterior density and MAP estimate

Choice of the prior: Dirichlet distribution

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

Dirichlet is the conjugate choice for multinomial

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

Posterior density

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

MAP estimate:

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

Dirichlet distribution:

\[ Dir(\theta | \alpha_1, \ldots, \alpha_k) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \cdots \theta_k^{\alpha_k-1} \]

Assume: \( k=3 \)

\[ \begin{aligned}
\alpha_k &= 10^{-1} \\
\alpha_k &= 10^1 
\end{aligned} \]

Expected value

The result is analogous to the result for binomial

\[ E(\theta) = \left[ 0 \leq \theta_i \leq 1, \sum_{i=1}^{k} \theta_i = 1 \right] Dir(\theta | \eta) d\theta = \left( \frac{\eta_1}{\eta_1 + \eta_2 + \eta_k}, \ldots, \frac{\eta_k}{\eta_1 + \eta_2 + \eta_k} \right) \]

Expectation based parameter estimate

\[ E(\theta) = \left( \frac{\alpha_i + N_i}{\alpha_i + N_i + \cdots + \alpha_k + N_k}, \ldots, \frac{\alpha_k + N_k}{\alpha_i + N_i + \cdots + \alpha_k + N_k} \right) \]

Represents the predictive probability of an event \( x=i \)

\[ P(x=i | \theta, \xi) = \frac{\alpha_i + N_i}{\alpha_i + N_i + \cdots + \alpha_k + N_k} \]
Other distributions

The same ideas can be applied to other distributions
- Typically we choose distributions that behave well so that computations lead to a nice solutions

• Exponential family of distributions

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)

Gamma distribution

• Gamma distribution

\[ \text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \]

\[ \mathbb{E}[\lambda] = \frac{a}{b} \quad \text{var}[\lambda] = \frac{a}{b^2} \]
Other distributions

**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:**

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

---

Gaussian (normal) distribution

- **Gaussian:** $x \sim \mathcal{N}(\mu, \sigma)$
- **Parameters:**
  - $\mu$ - mean
  - $\sigma$ - 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:**

![Gaussian Distribution](image)
Parameter estimates

- Loglikelihood
  \[ l(D, \mu, \sigma) = \log \prod_{i=1}^{n} p(x_i | \mu, \sigma) \]

- ML estimates of the mean and variance:
  \[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \]
  - 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:
  \[ \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \]

Multivariate normal distribution

- Multivariate normal:  \[ x \sim N(\mu, \Sigma) \]
- Parameters:  \[ \mu - \text{mean} \quad \Sigma - \text{covariance matrix} \]
- Density function:
  \[ p(x | \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \]
- Example:
Partitioned Gaussian Distributions

• Multivariate Gaussian:
  \[ p(x) = \mathcal{N}(x|\mu, \Sigma) \]
  – what are marginals and conditionals?

• Example:
  \[
  x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}
  \]
  \[
  \Lambda \equiv \Sigma^{-1} \quad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}
  \]
  Precision matrix

Partitioned Conditionals and Marginals

• Conditional density:
  \[
  p(x_a | x_b) = \mathcal{N}(x_a | \mu_{a|b}, \Sigma_{a|b})
  \]
  \[
  \Sigma_{a|b} = \Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba}
  \]
  \[
  \mu_{a|b} = \Sigma_{a|b} \{ \Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b) \}
  \]
  \[
  = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b)
  \]
  \[
  = \mu_a + \Sigma_{ab} \Sigma_{bb}^{-1} (x_b - \mu_b)
  \]

• Marginal Density:
  \[
  p(x_a) = \int p(x_a, x_b) \, dx_b
  \]
  \[
  = \mathcal{N}(x_a | \mu_a, \Sigma_{aa})
  \]
Partitioned Conditionals and Marginals

Parameter estimates

- Loglikelihood
  \[ l(D, \mu, \Sigma) = \log \prod_{i=1}^{n} p(x_i | \mu, \Sigma) \]

- ML estimates of the mean and covariances:
  \[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \]
  - Covariance estimate is biased
    \[ E_n(\hat{\Sigma}) = E_n\left( \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \right) = \frac{n-1}{n} \Sigma \neq \Sigma \]
  - Unbiased estimate:
    \[ \hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \]
Posterior of a multivariate normal

• Assume a prior on the mean \( \mu \) that is normally distributed:
  \[
p(\mu) \approx N(\mu_p, \Sigma_p)
  \]

• Then the posterior of \( \mu \) is normally distributed
  \[
p(\mu \mid D) \approx \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right]
  \]

  \[
  \approx \frac{1}{(2\pi)^{d/2} |\Sigma_p|^{1/2}} \exp \left[ -\frac{1}{2} (\mu - \mu_p)^T \Sigma_p^{-1} (\mu - \mu_p) \right]
  \]
  \[
  = \frac{1}{(2\pi)^{d/2} |\Sigma_n|^{1/2}} \exp \left[ -\frac{1}{2} (\mu - \mu_n)^T \Sigma_n^{-1} (\mu - \mu_n) \right]
  \]
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

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

- 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: A “new” prior

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