CS 2750 Machine Learning
Lecture 5

Density estimation II.

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

Outline

Outline:

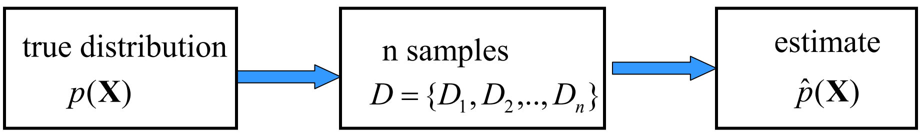
- **Density estimation:**
  - **Binomial distribution**
  - **Multinomial distribution**
  - **Normal distribution**
  - **Exponential family**

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 $p(\mathbf{X})$)**

Bernoulli trials

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 $(1-\theta)$

Probability of an outcome of a coin flip

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

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

N_1, N_2 - Number of heads and tails respectively

Posterior distribution

Posterior density

Likelihood of data → **prior**

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

Normalizing factor

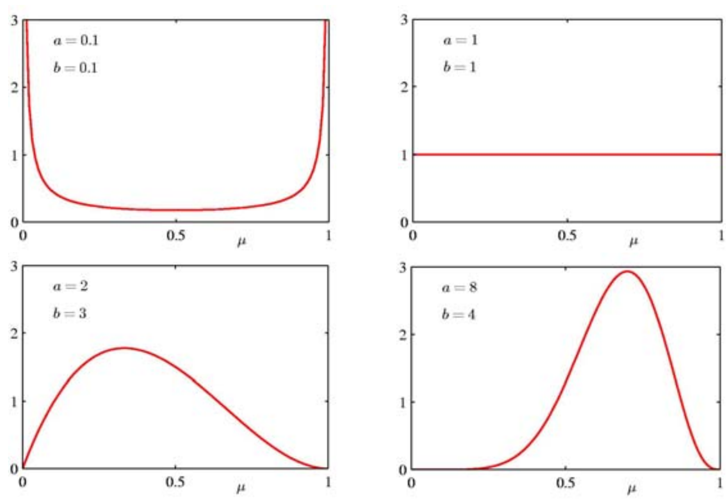
$$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)} = \theta^{N_1} (1-\theta)^{N_2}$$

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

Conjugate choice of prior: Beta

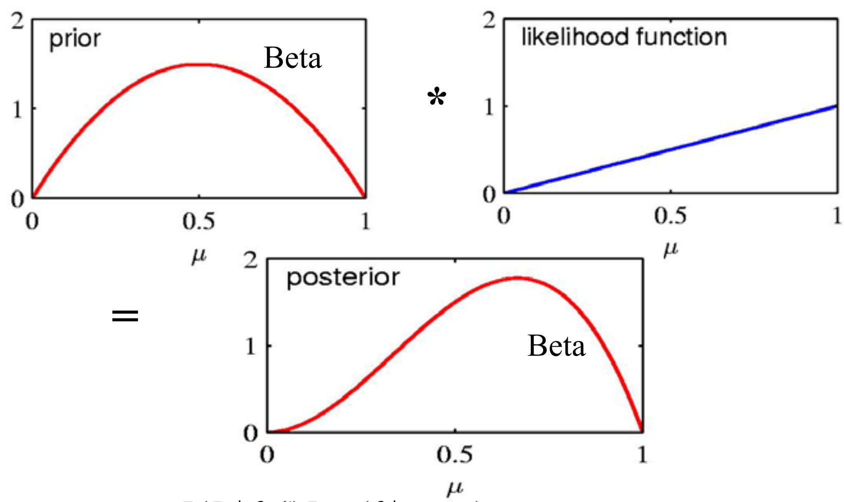
$$p(\theta \mid \xi) = Beta(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1} (1-\theta)^{\alpha_2 - 1}$$

Beta distribution



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

Posterior distribution



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

Bayesian framework

The ML estimate picks 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 $p(\theta \mid D,\xi) \approx Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$

- **The posterior can be used to define** $p(A \mid D)$**:**

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

Bayesian framework

- **Example: A probability of outcome $x=1$ in the next trial**
  $P(x=1|D,\xi)$

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

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

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

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

Expected value of the parameter

How to obtain the expected value?

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

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

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

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

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

Expected value 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:

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

Maximum a posterior probability

Maximum a posteriori estimate

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

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

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

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

Notice that parameters of the prior act like counts of heads and tails (sometimes they are also referred to as **prior counts**)

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

Binomial distribution

Example: 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 $(1-\theta)$

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

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

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

Multinomial distribution

Example: **Multi-way coin toss, or a roll of a dice**

- **Data:** a set of N trials (treated as a multi-set)

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

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

θ_i - probability of an outcome i

Probability of data (likelihood)

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

Multinomial distribution

ML estimate:

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

Posterior density and MAP estimate

Choice of the prior: **Dirichlet distribution**

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

Dirichlet is the conjugate choice for the multinomial

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

Posterior density

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

MAP estimate:

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

Expected value

The result is analogous to the result for the binomial

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

Expectation based parameter estimate

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

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 well so that 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 – Inverse Gamma**
  - **Gaussian - Gaussian (mean) and Wishart (covariance)**

Other distributions

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:

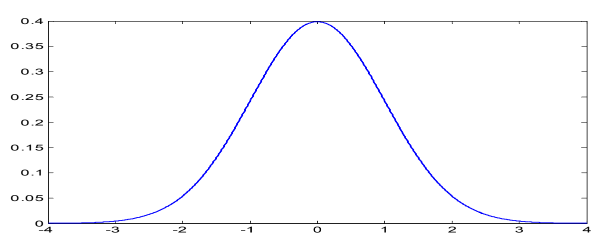
$$p(x \mid \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \qquad \text{for} \quad x \in \{0,1,2,\ldots\}$$

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

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

- **Unbiased estimate:**

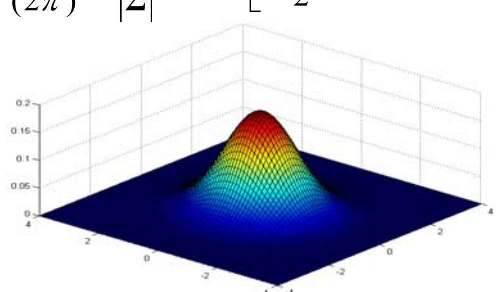
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Multivariate normal distribution

- **Multivariate normal:** $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **Parameters:** $\boldsymbol{\mu}$ - mean
  $\boldsymbol{\Sigma}$ - covariance matrix
- **Density function:**

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

- **Example:**



Parameter estimates

- **Loglikelihood** $l(D,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \log \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\mu},\boldsymbol{\Sigma})$
- **ML estimates of the mean and covariances:**

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

  – Covariance estimate is biased

$$E_n(\hat{\boldsymbol{\Sigma}}) = E_n\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T\right) = \frac{n-1}{n}\boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}$$

- **Unbiased estimate:**

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

Posterior of a multivariate normal

- **Assume a prior on the mean $\boldsymbol{\mu}$ that is normally distributed:**

$$p(\boldsymbol{\mu}) \approx N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$

- **Then the posterior of $\boldsymbol{\mu}$ is normally distributed**

$$p(\boldsymbol{\mu} \mid D) \approx \left(\prod_{i=1}^{n}\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right]\right)$$

$$* \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_p|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_p)^T\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_p)\right]$$

$$= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_n|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_n)\right]$$

Posterior of a multivariate normal

- **Then the posterior of $\boldsymbol{\mu}$ is normally distributed**

$$p(\boldsymbol{\mu} \mid D) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_n|^{1/2}} \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right]$$

$$\boldsymbol{\Sigma}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_p^{-1}$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_p \left(\boldsymbol{\Sigma}_p + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) + \frac{1}{n}\boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_p + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_p$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_p \left(\boldsymbol{\Sigma}_p + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \frac{1}{n}\boldsymbol{\Sigma}$$

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

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

- 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

$$p(D \mid \boldsymbol{\Theta}) = P(x \mid \boldsymbol{\Theta}) P(D_{n-1} \mid \boldsymbol{\Theta})$$

- **Then:**

$$p(\Theta \mid D, \xi) = \frac{P(x \mid \boldsymbol{\Theta}) \overbrace{P(D_{n-1} \mid \boldsymbol{\Theta})\, p(\Theta \mid \xi)}^{\text{A "new" prior}}}{\int_{\Theta} P(x \mid \boldsymbol{\Theta}) P(D_{n-1} \mid \boldsymbol{\Theta})\, p(\Theta \mid \xi)\, d\Theta}$$