

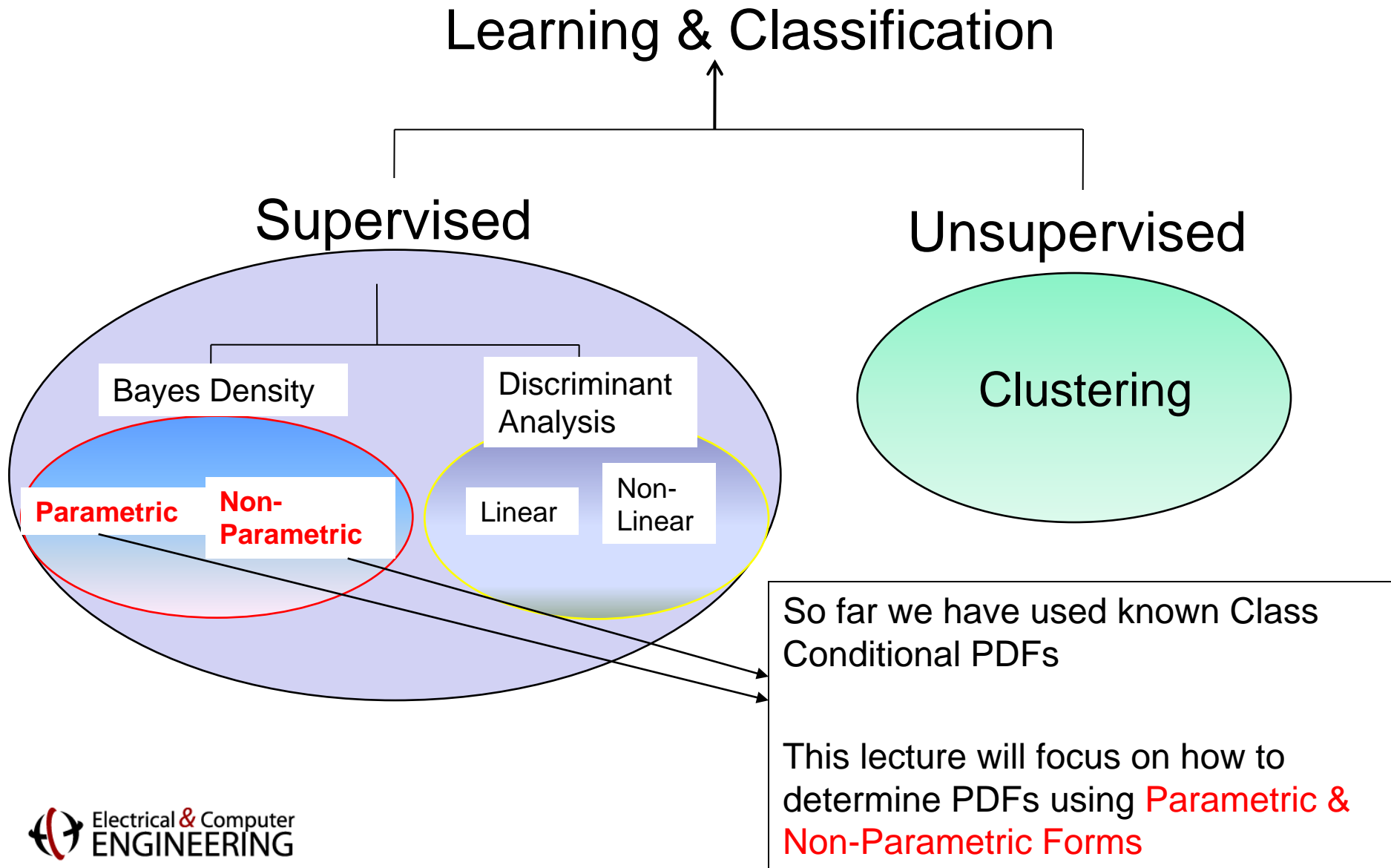
Prof. Marios Savvides

Pattern Recognition Theory

Lecture 5 : Parametric & Non-Parametric Density Estimation

All graphics from Pattern Classification, Duda, Hart and Stork, Copyright © John Wiley and Sons, 2001

Overview Of Pattern Recognition



Overview

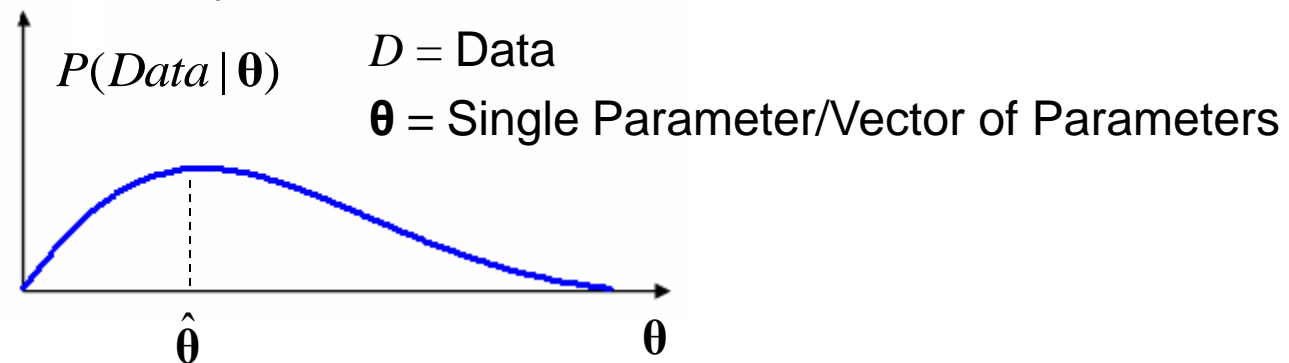
- Previous lectures have shown how to develop classifiers when the underlying statistical structure is known
- Parametric Estimation
 - This method assumes a **particular form** of a PDF (e.g. Gaussian) is known so that we only need to determine the **parameters** (e.g. Mean & Variance)
 - Maximum Likelihood Estimation (MLE)
 - Maximum A Posteriori (Bayesian) Estimation (MAPE)
- Non-Parametric Density Estimation
 - This method **does not assume ANY knowledge** about the density
 - Parzen Windows
 - Kernel Density Estimation
 - K-Nearest Neighbor Rule

ML Estimation (MLE)

- Maximum Likelihood Estimation

- Assume $P(\mathbf{x}|\omega)$ has a known parametric form uniquely determined by the parameter vector θ
- The parameters are assumed to be **FIXED (i.e. NON RANDOM)** but unknown
- Suppose we have a dataset D with the samples in D having been drawn **independently** according to the probability law $P(\mathbf{x}|\omega)$
- The MLE is the value of θ that best explains the data and **once we know this value, we know $P(\mathbf{x}|\omega)$**

$$\hat{\theta} = \arg \max_{\theta} \{P(D | \theta)\}$$



“Choose the value of θ that is the most likely to give rise to the data we observe”

MLE

$$D = \{x_1, x_2, \dots, x_N\} \quad \text{N independent observations}$$

$$P(D | \theta) = P(x_1, x_2, \dots, x_N | \theta) = \prod_{k=1}^N P(x_k | \theta)$$



The likelihood of observing
a particular pattern (random variable)

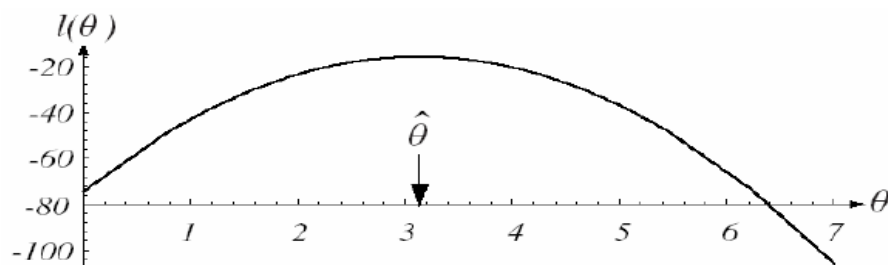
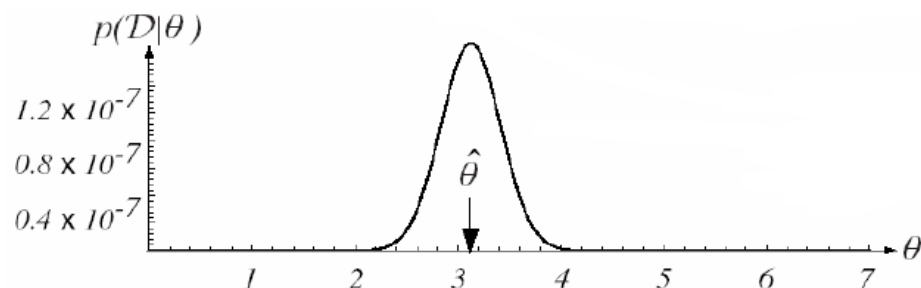
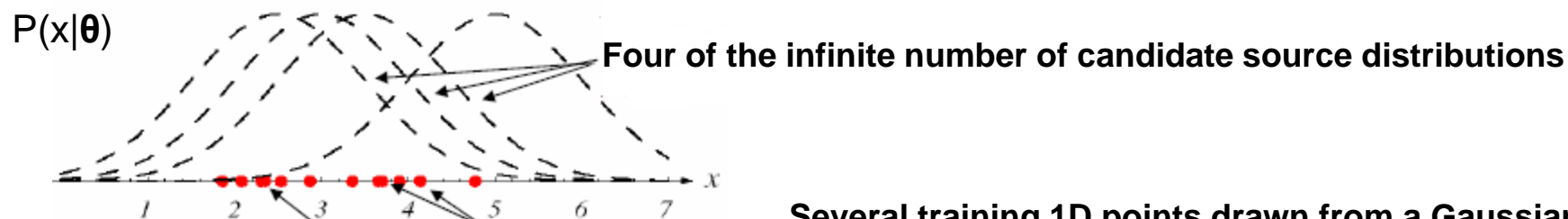
$$\hat{\theta} = \arg \max_{\theta} \{P(Data | \theta)\}$$

“Choose the value of θ that is the most likely to give rise to the data we observe”

MLE contd..

- It is convenient to work with the [log of the likelihood](#)

$$\hat{\theta} = \arg \max_{\theta} \{P(D|\theta)\} = \arg \max_{\theta} \{\log(P(D|\theta))\}$$



How To Solve For The ML Estimate?

- Let $\boldsymbol{\theta}$ be the p-component parameter vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_p]^T$
- Let this be the gradient operator $\nabla_{\boldsymbol{\theta}} = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \dots, \frac{\partial}{\partial \theta_p} \right]^T$
- We have $P(D | \boldsymbol{\theta}) = \prod_{k=1}^n P(x_k | \boldsymbol{\theta})$
- We define $l(\boldsymbol{\theta})$ the log-likelihood of the function

$$l(\boldsymbol{\theta}) = \log(P(D | \boldsymbol{\theta})) = \sum_{k=1}^n \log(P(x_k | \boldsymbol{\theta}))$$

- And

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log(P(D | \boldsymbol{\theta})) = \sum_{k=1}^n \nabla_{\boldsymbol{\theta}} \log(P(x_k | \boldsymbol{\theta}))$$

- A set of necessary condition for the ML estimate can be obtained from the set of p equations:

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \mathbf{0}$$

MLE Example: Univariate Gaussian

- Now assume **neither the mean nor the covariance** matrix are known
- First consider univariate case:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 = \mu \\ \theta_2 = \sigma^2 \end{bmatrix} \quad P(D | \boldsymbol{\theta}) = \prod_{k=1}^n P(x_k | \boldsymbol{\theta}) \quad \log P(x_k | \boldsymbol{\theta}) = -\frac{1}{2} \log(2\pi\theta_2) - \frac{1}{2\theta_2} (x_k - \theta_1)^2$$

- Its derivative is:

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \log(P(x_k | \boldsymbol{\theta})) = \begin{bmatrix} \frac{1}{\theta_2} (x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix} = 0$$

- Setting it to zero leads to:

$$\sum_{k=1}^n \frac{1}{\theta_2} (x_k - \theta_1) = 0 \quad \text{and} \quad -\sum_{k=1}^n \frac{1}{\theta_2} + \sum_{k=1}^n \frac{(x_k - \theta_1)^2}{\theta_2^2} = 0$$

- Rearranging:

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k$$

Sample Mean

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

Sample Variance

ML Estimate

MLE Example: Multivariate Gaussian

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

$$l(\theta) = \log(P(D | \theta)) = \sum_{k=1}^n \log(P(\mathbf{x}_k | \theta))$$

Consider **only the mean is unknown**:

$$\log P(\mathbf{x}_k | \boldsymbol{\mu}) = -\frac{1}{2} \log\left((2\pi)^d |\boldsymbol{\Sigma}|\right) - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

Derivative of log likelihood must be set to 0 to obtain the MLE

$$\nabla_{\boldsymbol{\mu}} \log(P(D | \boldsymbol{\mu})) = \sum_{k=1}^n \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) = \mathbf{0}$$

The ML estimate must satisfy:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$$

Sample Mean -> ML Estimate

MLE Example: Multivariate Gaussian

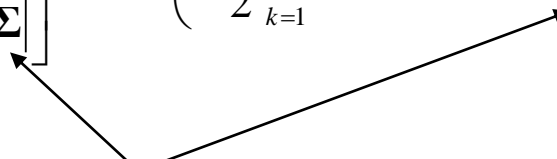
- Neither the mean nor the covariance matrix are known

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 = \boldsymbol{\mu} \\ \theta_2 = \boldsymbol{\Sigma} \end{bmatrix} \quad \log P(\mathbf{x}_k | \boldsymbol{\theta}) = -\frac{1}{2} \log \left((2\pi)^d |\boldsymbol{\Sigma}| \right) - \frac{1}{2} (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu})$$

- Derivative of log likelihood is:

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \log(P(\mathbf{x}_k | \boldsymbol{\theta})) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) \\ ? \end{bmatrix}$$

- How to take the gradient of a determinant of a matrix?

$$\begin{aligned} P(\mathbf{x} | \boldsymbol{\Sigma}) &= \prod_{k=1}^n P(\mathbf{x}_k | \boldsymbol{\Sigma}) = \prod_{k=1}^n \left\{ \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp \left(-\frac{1}{2} (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) \right) \right\} \\ &= \frac{1}{\left[(2\pi)^d |\boldsymbol{\Sigma}| \right]^{n/2}} \exp \left(-\frac{1}{2} \sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) \right) \end{aligned}$$


Need to take gradient with respect to $\boldsymbol{\Sigma}$

MLE Example: Multivariate Gaussian

$$P(\mathbf{x} | \Sigma) = \prod_{k=1}^n P(\mathbf{x}_k | \Sigma) = \prod_{k=1}^n \left\{ \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp \left(-\frac{1}{2} (\mathbf{x}_k - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) \right) \right\}$$

$$= \frac{1}{\left[(2\pi)^d |\Sigma| \right]^{n/2}} \exp \left(-\frac{1}{2} \sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) \right)$$

Scalar



$$\mathbf{b}^T \mathbf{B} \mathbf{b} = \text{trace}(\mathbf{b}^T \mathbf{B} \mathbf{b}) = \text{trace}(\mathbf{B} \mathbf{b} \mathbf{b}^T)$$

$$\text{trace}(\mathbf{A} + \mathbf{B}) = \text{trace}(\mathbf{A}) + \text{trace}(\mathbf{B})$$

$$\text{trace}(\mathbf{C}(\mathbf{A} + \mathbf{B})) = \text{trace}(\mathbf{C}\mathbf{A}) + \text{trace}(\mathbf{C}\mathbf{B})$$

$$\sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_k - \boldsymbol{\mu})$$

$$= (\mathbf{x}_1 - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}) + \dots + (\mathbf{x}_n - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu})$$

$$= \text{trace} \left(\Sigma^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}) (\mathbf{x}_1 - \boldsymbol{\mu})^T \right) + \dots + \text{trace} \left(\Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) (\mathbf{x}_n - \boldsymbol{\mu})^T \right)$$

$$= \text{trace} \left(\Sigma^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}) (\mathbf{x}_1 - \boldsymbol{\mu})^T + \dots + \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}) (\mathbf{x}_n - \boldsymbol{\mu})^T \right)$$

$$= \text{trace} \left(\Sigma^{-1} \sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu}) (\mathbf{x}_k - \boldsymbol{\mu})^T \right)$$



$$\mathbf{A} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu}) (\mathbf{x}_k - \boldsymbol{\mu})^T$$

MLE Example: Multivariate Gaussian

We can now rewrite:

$$\mathbf{A} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T$$

$$\sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}) = \text{trace} \left(\boldsymbol{\Sigma}^{-1} \sum_{k=1}^n (\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^T \right) = \text{trace} (n \boldsymbol{\Sigma}^{-1} \mathbf{A}) = n \cdot \text{trace} (\boldsymbol{\Sigma}^{-1} \mathbf{A})$$

$$\begin{aligned} p(\mathbf{x} | \boldsymbol{\Sigma}) &= \prod_{k=1}^n p(\mathbf{x}_k | \boldsymbol{\Sigma}) = \frac{1}{[(2\pi)^d |\boldsymbol{\Sigma}|]^{n/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \\ &= \frac{1}{[(2\pi)^{dn/2} |\boldsymbol{\Sigma}|]^{n/2}} \exp \left[-\frac{n}{2} \text{trace}(\boldsymbol{\Sigma}^{-1} \mathbf{A}) \right] \end{aligned}$$

- Now define $\mathbf{B} = \boldsymbol{\Sigma}^{-1} \mathbf{A}$ and let $\lambda_1, \lambda_2, \dots, \lambda_d$ be its eigenvalues

$$\begin{aligned} p(\mathbf{x} | \boldsymbol{\Sigma}) &= \frac{1}{[(2\pi)^{dn/2} |\boldsymbol{\Sigma}|]^{n/2}} \exp \left[-\frac{n}{2} \text{trace}(\mathbf{B}) \right] \\ &= \frac{1}{(2\pi)^{dn/2}} \left[\frac{|\mathbf{B}|}{|\mathbf{A}|} \right]^{n/2} \exp \left[-\frac{n}{2} \text{trace}(\mathbf{B}) \right] \\ &= \frac{1}{(2\pi)^{dn/2}} |\mathbf{A}|^{-n/2} \left(\prod_{i=1}^d \lambda_i \right)^{n/2} \exp \left[-\frac{n}{2} \sum_{i=1}^d \lambda_i \right] \end{aligned}$$

$$\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$$

$$\det(\mathbf{A}^{-1}) = \det(\mathbf{A})^{-1}$$

$$\text{trace}(\mathbf{A}) \quad \text{Sum of eigenvalues}$$

$$\det(\mathbf{A}) \quad \text{Product of eigenvalues}$$

MLE Example: Multivariate Gaussian

Now consider the log likelihood

$$P(\mathbf{x} | \mathbf{\Sigma}) = (2\pi)^{-dn/2} |\mathbf{A}|^{-n/2} \left(\prod_{i=1}^d \lambda_i \right)^{n/2} \exp \left(-\frac{n}{2} \sum_{i=1}^d \lambda_i \right)$$

$$\log P(\mathbf{x} | \mathbf{\Sigma}) = -\frac{dn}{2} \log(2\pi) - \frac{n}{2} \log |\mathbf{A}| + \frac{n}{2} \sum_{i=1}^d \log(\lambda_i) - \frac{n}{2} \sum_{i=1}^d \lambda_i$$

Remember that \mathbf{A} is fixed, and taking the gradient with respect to $\mathbf{\Sigma}$ is now equivalent to taking the derivatives with respect to the eigenvalues of \mathbf{B}

$$\frac{\partial}{\partial \lambda_i} \log P(\mathbf{x} | \mathbf{\Sigma}) = \frac{n}{2\lambda_i} - \frac{n}{2} = 0 \Rightarrow \lambda_i = 1 \text{ for } i = 1, \dots, d$$

Thus \mathbf{B} must have all eigenvalues of 1 and is equivalent to \mathbf{I} . This means that the likelihood is maximized if $\mathbf{\Sigma}^{-1}\mathbf{A}=\mathbf{I}$ or $\mathbf{\Sigma}=\mathbf{A}$

$$\hat{\mathbf{\Sigma}}_{\text{ML}} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \hat{\mathbf{\mu}})(\mathbf{x}_k - \hat{\mathbf{\mu}})^T$$

MLE - Sample Covariance

MLE vs The Bayesian Approach

- **Maximum Likelihood Estimation (MLE)**

- The parameters are assumed to be **FIXED** (i.e. **NON RANDOM**) but unknown
- The ML seeks the solutions that best explains the data

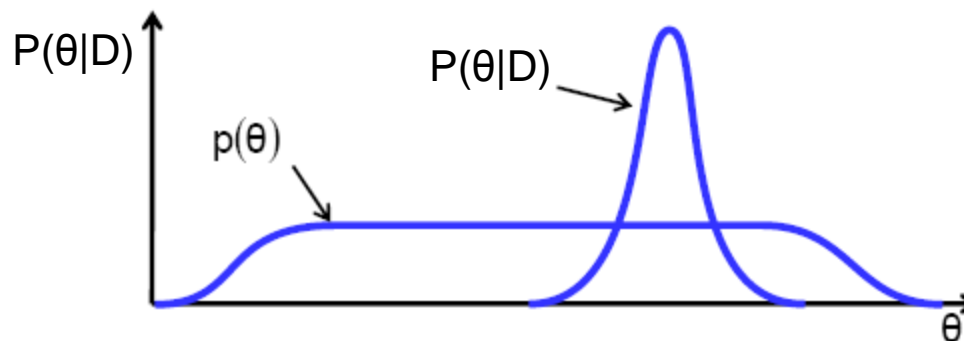
$$\hat{\theta} = \arg \max_{\theta} \{P(Data | \theta)\}$$

- **Bayesian Estimation (BE)**

- The parameters are assumed to be **RANDOM VARIABLES** with some known **PRIORI DISTRIBUTION**
- Bayesian approach aims at estimating the posterior density $P(\theta | Data)$
- **The MAPE (Maximum A Posteriori Estimate)** of θ is the value of θ that maximizes the posterior density (i.e. it is the mode of the posterior)

Bayesian Estimation

- In the Bayesian approach, our uncertainty about the parameters is represented by a PDF
- The parameters are described by a prior density $P(\theta)$ which indicates which parameters are more likely than others
- We make use of Bayes theorem to find the posterior $P(\theta|D)$
- Ideally, we want the training data to “sharpen” the posterior $P(\theta|D)$ or reduce our uncertainty about the parameters



Bayesian Estimation

- We ideally want to estimate a PDF - $P(x)$. Best we can do is estimate it by observing the training data to obtain $P(x|D)$. We also assume that it has a known parametric form. So $P(x|\theta)$ is completely known. But θ is random (unlike in MLE) and has its own PDF.

$$P(x|D) = \int P(x, \theta | D) d\theta \quad \leftarrow \text{Using the theorem of total probability}$$

$$= \int P(x|\theta) P(\theta|D) d\theta$$

Known

Unknown

This integration is typically very hard

θ is random and has its own PDF

Applying Bayes rule: $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$

Posterior

$$P(D|\theta) = \prod_{k=1}^N P(x_k|\theta)$$

MAP Estimation (MAPE)

- MLE aims at determining the value $\theta = \theta_{\text{MLE}}$ that maximizes the likelihood

$$\theta_{\text{MLE}} = \arg \max_{\theta} \{P(D | \theta)\}$$

- MAPE assumes that θ is not a fixed, but is an RV with a PDF given by $P(\theta)$ and it aims at maximizing the posterior

$$\theta_{\text{MAP}} = \arg \max_{\theta} \{P(\theta | D)\}$$

$$\text{where } P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)} = \frac{P(D | \theta)P(\theta)}{\int P(D | \theta)P(\theta)d\theta} = \frac{P(\theta) \prod_{k=1}^N P(x_k | \theta)}{\int P(D | \theta)P(\theta)d\theta}$$

Since $P(D)$ is constant we can also view MAP estimate as

$$\theta_{\text{MAP}} = \arg \max_{\theta} \{P(D | \theta)P(\theta)\}$$

- Thus the MAPE of θ is simply the mode of the posterior $P(\theta | D)$ and MAPE differs from MLE as it determines a value of θ which maximizes the posterior instead of the likelihood

MAPE Example: Univariate Gaussian

- Compute $P(\theta|D)$ and the desired PDF $P(x|D)$ where $p(\mu) = N(\mu_0, \sigma_0^2)$
- Assume the known prior knowledge about the mean can be expressed by a *known* prior density assumed to be normal:

$$p(\mu) = N(\mu_0, \sigma_0^2) \quad \text{Known!}$$

$$P(\mu | D) = \frac{P(D | \mu)P(\mu)}{\int P(D | \mu)P(\mu)d\mu} = \alpha \prod_{k=1}^n P(x_k | \mu)P(\mu)$$

$$\begin{aligned}
 P(\mu | D) &= \alpha \prod_{k=1}^n \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma}\right)^2\right]}^{P(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right]}^{P(\mu)} \\
 &= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^n \left(\frac{x_k - \mu}{\sigma}\right)^2 + \left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right)\right] \\
 &= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2} \sum_{k=1}^n x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right] \quad \text{GAUSSIAN!}
 \end{aligned}$$

MAPE Example: Univariate Gaussian

Gaussian Likelihood && Gaussian Prior \rightarrow Gaussian Posterior

compare

$$P(\mu | D) = \alpha \exp \left[-\frac{1}{2} \left[\underbrace{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)}_{\text{precision}} \mu^2 - 2 \underbrace{\left(\frac{1}{\sigma^2} \sum_{k=1}^n x_k + \frac{\mu_0}{\sigma_0^2} \right)}_{\text{precision} \times \text{mean}} \mu \right] \right]$$

$$P(\mu | D) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \left[-\frac{1}{2} \left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right] = \alpha \exp \left[-\frac{1}{2} \left[\mu^2 \left(\frac{1}{\sigma_n^2} \right) - 2\mu \left(\frac{\mu_n}{\sigma_n^2} \right) + \left(\frac{\mu_n}{\sigma_n} \right)^2 \right] \right]$$

We get $\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$ and $\frac{\mu_n}{\sigma_n^2} = \frac{1}{\sigma^2} \sum_{k=1}^n x_k + \frac{\mu_0}{\sigma_0^2} = \frac{n}{\sigma^2} \hat{\mu}_n + \frac{\mu_0}{\sigma_0^2}$

where $\hat{\mu}_n = \frac{1}{n} \sum_{k=1}^n x_k$ and is the sample mean

So

$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \right) \mu_0$$

MAPE Example: Univariate Gaussian

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \right) \mu_0$$

$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

Best guess for μ

Uncertainty

- μ_n is a linear combination of $\hat{\mu}_n$ and μ_0 (always lies between them)
 - For small samples -> More weights on prior ($\hat{\mu}_0$)
 - For large samples -> More weights on observation ($\hat{\mu}_n$)
- If $\sigma_n \neq 0$, then $\mu_n \rightarrow \hat{\mu}_n$ as $n \rightarrow \infty$ (MLE = MAPE)
- If $\sigma_0 = 0$, then $\mu_n = \mu_0$
- If $\sigma_0 \gg \sigma$ then $\mu_n = \hat{\mu}_n$ (MLE = MAPE)

MAPE Example: Univariate Gaussian

- μ_n represents our best guess for μ after observing n samples, and σ_n^2 measures our uncertainty about this guess.

$$\mu_n = \underbrace{\left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \right)}_{ML} \hat{\mu}_n + \underbrace{\left(\frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \right)}_{PRIOR_KNOWLEDGE} \mu_0$$

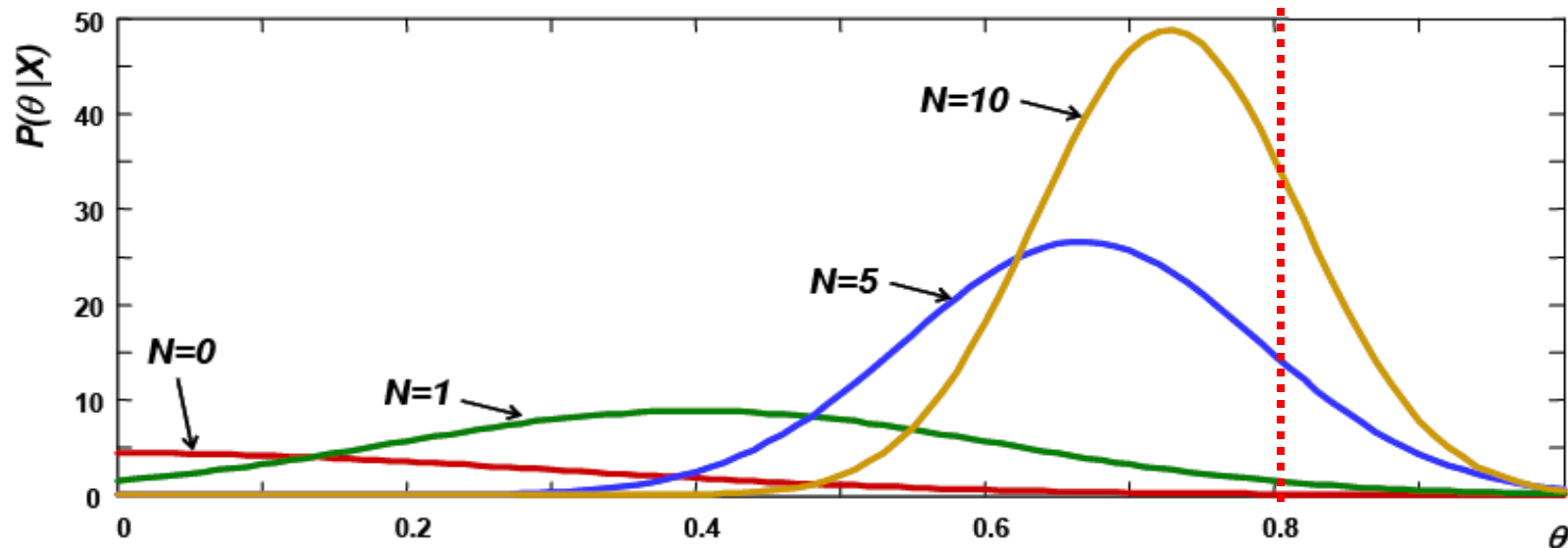
$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

Decreases as $n \rightarrow \infty$

- Each additional observation decreases our uncertainty and $P(\mu|D)$ becomes narrower and sharply peaked around the true value of μ and becomes a Dirac delta function. This is called **Bayesian Learning**.
- A class of PDF $P(\theta)$ is said to be **conjugate** to a class of likelihood functions $P(x|\theta)$ if the resulting posterior distributions $P(\theta|x)$ are in the **same family as $P(\theta)$** .
 - The Gaussian family is conjugate to itself if the likelihood function is Gaussian, choosing a Gaussian prior will ensure that the posterior distribution is also Gaussian

MAPE Example: Gaussians

- Assume that the true mean of the $P(x)$ is $N(0.8, 0.09)$.
 - In reality this is something we cannot know
- We generate a number of examples from this distribution
- We don't know where the mean will be, so we assume a Gaussian prior
 - $P_0(\mu) = N(0, 0.09)$
- As the number of training examples increases, the estimates μ_N approaches its true value of 0.8 and the spread decreases



MAPE: Back To The Gaussian Example

So far we've only figured $P(\mu|D)$. We want $P(x|D)$

$$P(x|D) = \int P(x|\mu)P(\mu|D)d\mu$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu$$

$$= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2} \frac{(x-\mu_n)^2}{\sigma^2 + \sigma_n^2}\right] f(\sigma, \sigma_n)$$

Not dependent on μ !!

$$f(\sigma, \sigma_n) = \int \exp\left[-\frac{1}{2} \frac{\sigma^2 + \sigma_n^2}{\sigma^2 \sigma_n^2} \left(\mu - \frac{\sigma_n^2 x + \sigma^2 \mu_n}{\sigma^2 + \sigma_n^2}\right)^2\right] d\mu$$

$$P(x|D) = N(\mu_n, \sigma^2 + \sigma_n^2)$$

Thus $P(x|D)$ is the desired class-conditional density $P(x|\omega_i, D)$ and together with the priors $P(\omega_i)$ it gives us the probabilistic framework needed to design the classifier. This is in contrast with the ML method that only makes **point estimates** rather than estimate a **distribution** for $P(x|D)$

MAPE Example: Multivariate Gaussian

- Compute $P(\boldsymbol{\mu}|D)$ and the desired PDF $P(\mathbf{x}|D)$ where $P(\mathbf{x}|\boldsymbol{\mu}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Assume the known prior knowledge about the mean can be expressed by a *known* prior density assumed to be normal:

$$P(\boldsymbol{\mu}) = N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \quad \text{Known!}$$

$$P(\boldsymbol{\mu} | D) = \frac{P(D | \boldsymbol{\mu})P(\boldsymbol{\mu})}{\int P(D | \boldsymbol{\mu})P(\boldsymbol{\mu})d\boldsymbol{\mu}} = \alpha \prod_{k=1}^n P(\mathbf{x}_k | \boldsymbol{\mu})P(\boldsymbol{\mu})$$

$$P(\boldsymbol{\mu} | D) = \alpha \prod_{k=1}^n P(\mathbf{x}_k | \boldsymbol{\mu})P(\boldsymbol{\mu})$$

$$= \alpha \exp \left[-\frac{1}{2} \left(\boldsymbol{\mu}^T (n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}) \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \left(\boldsymbol{\Sigma}^{-1} \sum_{k=1}^n \mathbf{x}_k + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right) \right) \right]$$

Which has a Gaussian Form

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

MAPE Example: Multivariate Gaussian

- Thus $P(\boldsymbol{\mu}|D)$ is $N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Equating coefficients, we obtain:

$$\boldsymbol{\Sigma}_n^{-1} = n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1} \quad \text{and} \quad \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n = n\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_n + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$$

where $\hat{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$ is the sample mean

After some manipulation (which we don't prove) we get:

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \hat{\boldsymbol{\mu}}_n + \frac{1}{n} \boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0 + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_0$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{n} \boldsymbol{\Sigma} \right)^{-1} \frac{1}{n} \boldsymbol{\Sigma}$$

- Finally, it can be shown that

$$P(\mathbf{x} | D) = N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$$

MAPE Example: Gaussians

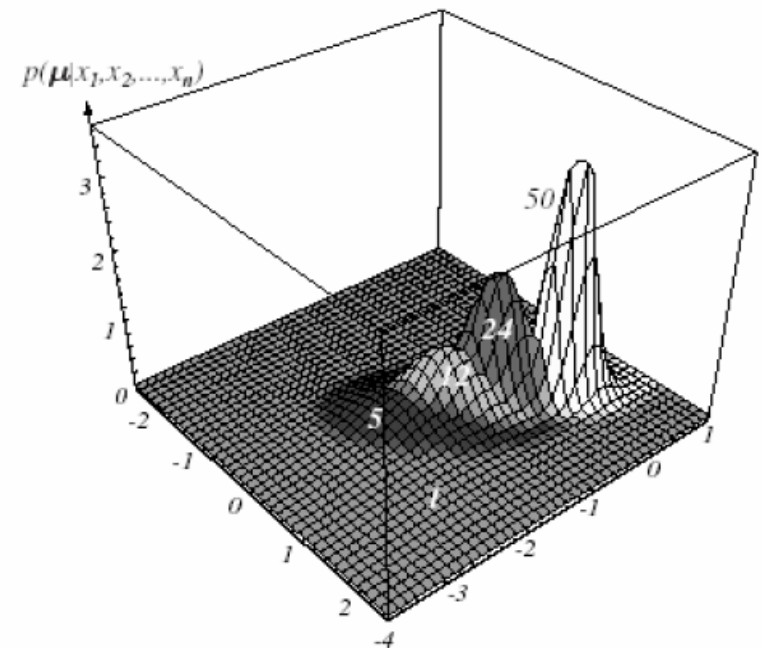
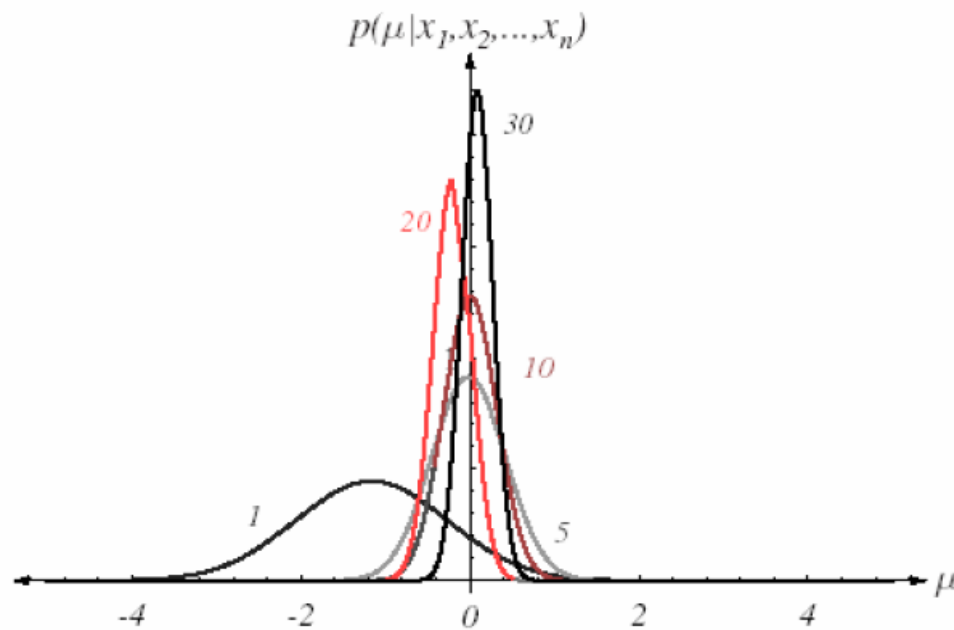


FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Recap: ML vs Bayesian Estimation

- BE assumes that the parameters come from a distribution with known priors
- BE provides a distribution for θ rather than point values. BE provides more information, but is in general much more difficult to compute (integration)
- For most times, if the assumptions are correct, MLE gives good enough results

Recap: ML vs Bayesian Estimation

- Amount of Training data
 - The two methods are equivalent assuming infinite training samples.
 - They differ for smaller training data
- Computational Complexity
 - ML uses **differential calculus** or gradient search
 - Bayesian estimation needs complex multidimensional **integration techniques**
- Solution Complexity
 - ML solution is easy to interpret
 - A Bayes Estimation solution **might not be** of the parametric form assumed
- Prior distribution
 - If the prior $P(\theta)$ is **uniform (flat)**, BE solutions are equivalent to ML solutions

Conjugate Priors

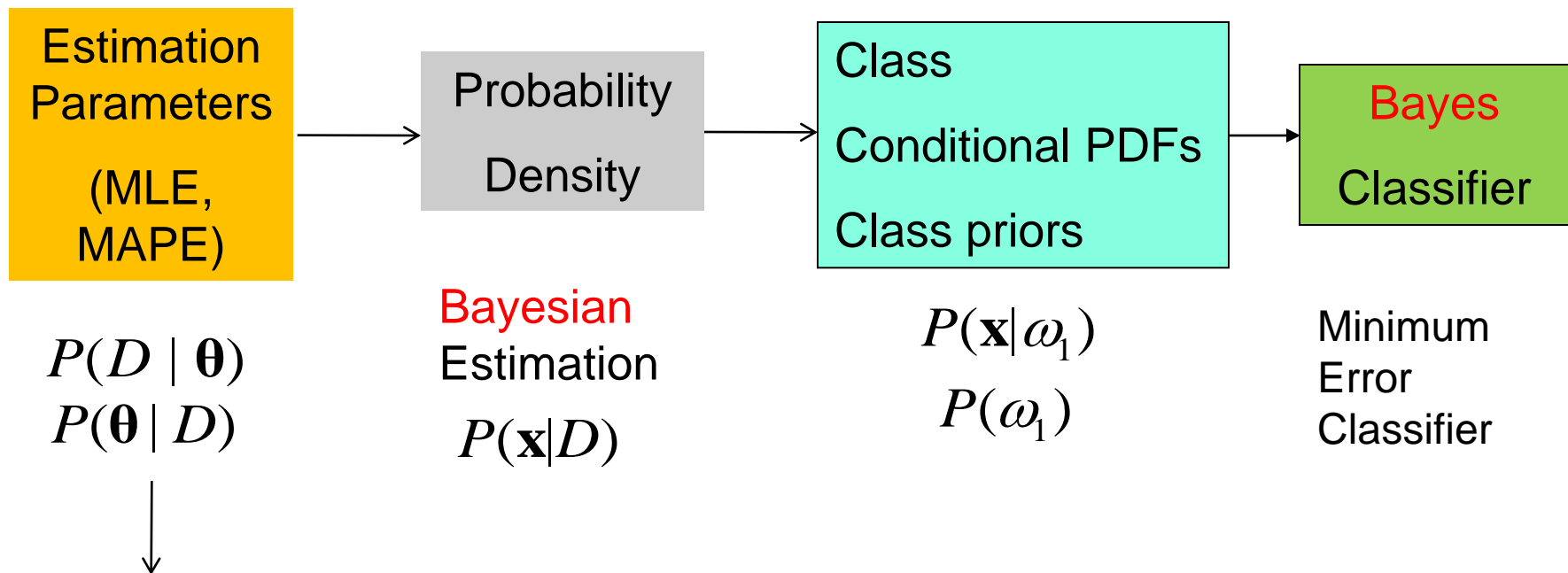
Discrete likelihood distributions

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters
Bernoulli	p (probability)	Beta	α, β ^[4]	$\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i$
Binomial	p (probability)	Beta	α, β ^[4]	$\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n N_i - x_i$
Poisson	λ (rate)	Gamma	α, β ^[4]	$\alpha + \sum_{i=1}^n x_i, \beta + n$
Multinomial	\mathbf{p} (probability vector)	Dirichlet	$\vec{\alpha}$	$\vec{\alpha} + \sum_{i=1}^n \vec{x}^{(i)}$
Geometric	p_0 (probability)	Beta	α, β ^[4]	$\alpha + n, \beta + \sum_{i=1}^n x_i$

Continuous likelihood distributions

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters
Uniform	$U(0, \theta)$	Pareto	x_m, k	$\max\{x_{(n)}, x_m\}, k + n$
Exponential	λ (rate)	Gamma	α, β ^[4]	$\alpha + n, \beta + \sum_{i=1}^n x_i$
Normal with known variance σ^2	μ (mean)	Normal	μ_0, σ_0^2	$(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2}) / (\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}), (\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2})^{-1}$
Normal with known precision τ	μ (mean)	Normal	μ_0, τ_0	$(\tau_0 \mu_0 + \tau \sum_{i=1}^n x_i) / (\tau_0 + n\tau), \tau_0 + n\tau$

Big Picture



Recap

- Maximum Likelihood Estimation (MLE)
 - MLE for Univariate and Multivariate Gaussians
- Bayesian Estimation (BE)
 - BE for Univariate and Multivariate Gaussians
- Maximum A Posteriori Estimation (MAPE)
- Conjugate Priors