Machine learning

Parametric Models
Part I: Maximum Likelihood and Bayesian Density Estimation

Mohammad-Reza A. Dehaqani

dehaqani@ut.ac.ir

Introduction

- Bayesian Decision Theory shows us how to design an **optimal classifier** if we know the **prior probabilities** $P(\omega_i)$ and the **class-conditional densities** $p(x|\omega_i)$.
- Unfortunately, we **rarely have complete** knowledge of the **probabilistic structure**.
- However, we can often find **design samples** or **training data** that include **particular representatives of the patterns** we want to classify.

Introduction

- To **simplify** the problem, we can assume some **parametric form** for the conditional densities and **estimate** these parameters using training data.
- Then, we can **use** the resulting **estimates** as if they were the true values and **perform classification** using the Bayesian decision rule.
- We will consider only the **supervised learning** case where the true class label for each sample is known.

Maximum likelihood vs. Bayesian

- We will study two **estimation** procedures:
  - **Maximum likelihood** estimation
    - Views the **parameters as quantities** whose values are **fixed** but **unknown**.
    - Estimates these values by maximizing the probability of obtaining the samples observed.
  - **Bayesian estimation**
    - Views the **parameters as random variables** having some **known prior distribution**.
    - Observing new samples **converts the prior to a posterior** density.

Maximum Likelihood Function

- Suppose we have a set $\boldsymbol{D} = \{x_1, \ldots, x_n\}$ of **independent and identically distributed** (**i.i.d.**) samples drawn from the **density** $p(x|\theta)$.
- We would like to use training samples in $\boldsymbol{D}$ to **estimate** the unknown **parameter** vector θ.
- Define $L(\theta|\boldsymbol{D})$ as the **likelihood function** of θ with respect to $\boldsymbol{D}$ as

$$L(\boldsymbol{\theta}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{x_1}, \ldots, \mathbf{x_n}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x_i}|\boldsymbol{\theta}).$$

Maximum Likelihood Estimation

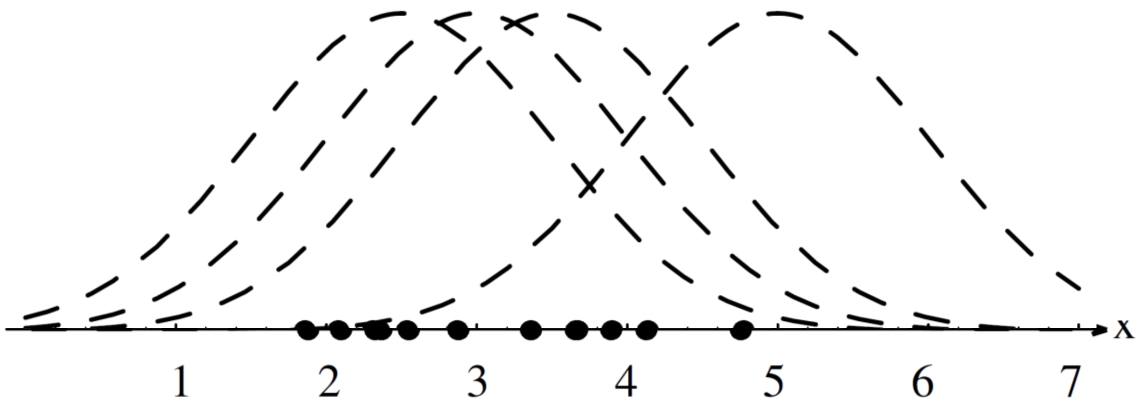
- The maximum likelihood estimate (**MLE**) of θ is, by definition, the value $\hat{\theta}$ **that maximizes** $L(\theta|\boldsymbol{D})$ and can be computed as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathcal{D}).$$

- It is often easier to work with the **logarithm** of the likelihood function (log-likelihood function) that gives

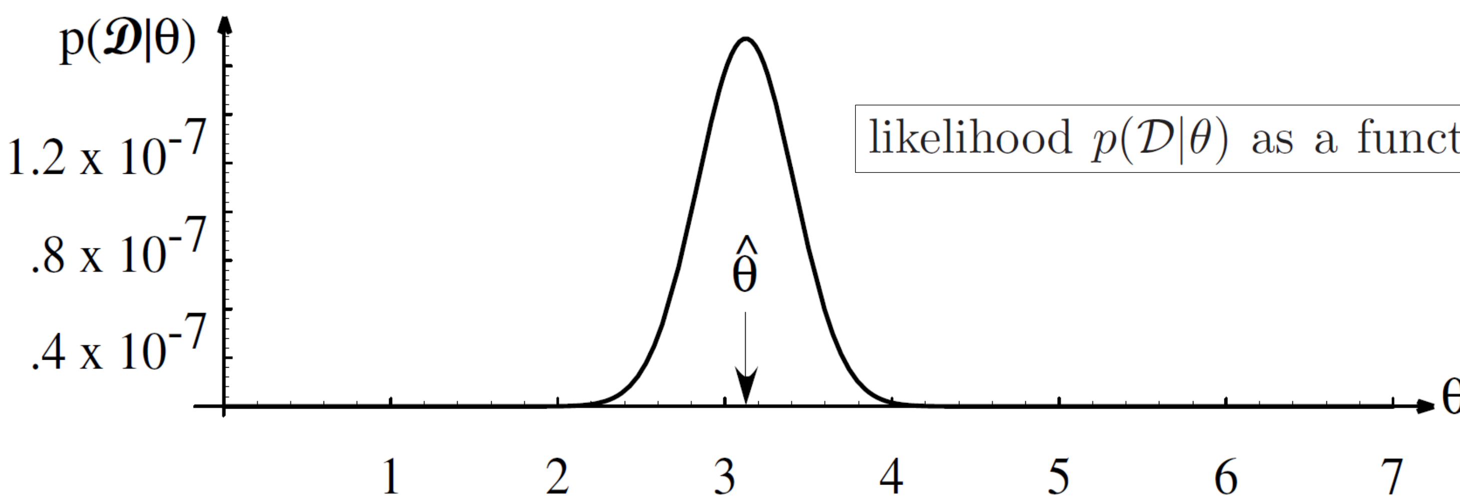
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\mathcal{D}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta}).$$

Example



Some of candidate source distributions

Several training points assumed to be drawn from a Gaussian



likelihood $p(\mathcal{D}|\theta)$ as a function of the mean.

the likelihood lies in a different space from $p(x|\hat{\theta})$.

Maximum Likelihood Estimation

- If the number of parameters is p, i.e., $\theta = (\theta_1, \theta_2. . . , \theta_p)^T$, define the **gradient operator**

$$\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\theta_1}} \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{\theta_p}} \end{bmatrix}.$$

- Then, the **MLE** of θ should satisfy the necessary conditions

$$\nabla_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\mathcal{D}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}_{\boldsymbol{i}}|\boldsymbol{\theta}) = 0.$$

Properties of MLEs

- The MLE is the **parameter point** for which the **observed sample is the most likely**.

- The procedure with partial derivatives may result in **several local extrema**. We should check each solution individually to identify the **global optimum**.

- **Boundary conditions** must also be checked separately for extrema.

- **Invariance** property: if $\hat{\theta}$ is the MLE of θ, then for any function $f(\theta)$, the MLE of $f(\theta)$ is $f(\hat{\theta})$.

The Gaussian Case

- Suppose that $p(x|\theta) = N(\mu,\Sigma)$.
  - When Σ is **known** but μ is **unknown**

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

  - When **both** μ and Σ are **unknown**

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

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\boxed{\theta_1 = \mu \text{ and } \theta_2 = \sigma^2.}$$

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

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0$$

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0,$$

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

Analysis of the multivariate case $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t.$

The Bernoulli Case

$$L(p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{(1-x_i)}$$

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

$$\frac{\partial \ell(p)}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\sum_{i=1}^{n} (1-x_i)}{1-p} \overset{\text{set}}{=} 0$$

$$\sum_{i=1}^{n} x_i - p \sum_{i=1}^{n} x_i = p \sum_{i=1}^{n} (1-x_i)$$

$$p = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Bias of Estimators

- Bias of an estimator $\hat{\theta}$ is the **difference between the expected value of $\hat{\theta}$ and θ**.
- The MLE of μ is an **unbiased** estimator for μ because $E[\hat{\mu}] = \mu$
- The MLE of Σ is **not an unbiased** estimator for because $E[\hat{\Sigma}] = \frac{n-1}{n}\Sigma \neq \Sigma$.
- The **sample covariance**

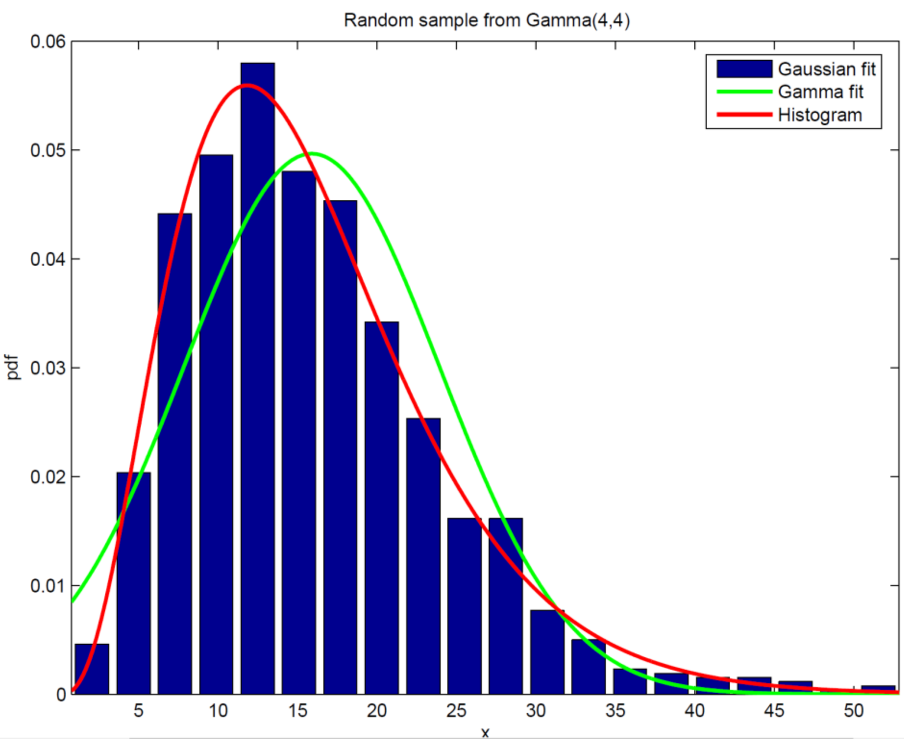
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

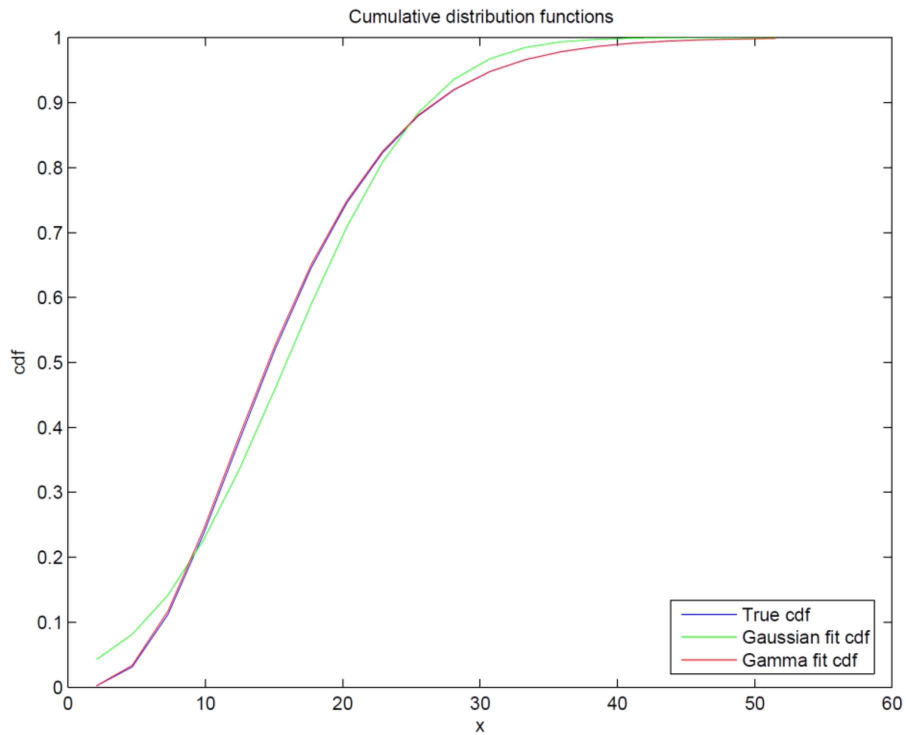
is an unbiased estimator for Σ

Goodness-of-fit

- To measure **how well a fitted distribution** resembles the **sample data** (goodness-of-fit), we can use the **Kolmogorov-Smirnov** test statistic.
  - It is defined as the **maximum value** of the absolute difference between the **cumulative distribution function** estimated from the sample and the one calculated from the fitted distribution.
- After estimating the parameters for different distributions, we can compute the Kolmogorov-Smirnov statistic for each distribution and choose the one with the **smallest value** as the **best fit to** our sample.

Estimated pdf





True pdf is Gamma(4, 4). Estimated pdfs are N(15.8, 62.1) and Gamma(4.0, 3.9).

Cumulative distribution functions for the example

Bayesian Estimation

- **Bayesian estimation or Bayesian learning** approach to pattern classification problems.
- The method is nearly identical to maximum likelihood, there is **a conceptual difference**:
  - **In ML $\boldsymbol{\theta}$**, to be fixed, in Bayesian learning we consider $\boldsymbol{\theta}$ to be a **random variable**,
  - **Training data** allows us to **convert** a distribution on this variable **into a posterior probability density.**
- Suppose the set $\boldsymbol{D} = \{x_1, \ldots, x_n\}$ contains the samples drawn **independently** from the **density** $p(x|\theta)$ whose **form is assumed** to be **known** but θ is **not known exactly**.
- Assume that θ is a **quantity** whose **variation** <u>can be described by</u> the **prior probability distribution** $p(\theta)$.

The Class-Conditional Densities

- We compute $P(\omega_i|\mathbf{x})$ using **all of the information** at our disposal
- Given the sample ***D***, Bayes' formula then becomes

$$\boxed{P(\omega_i|\mathbf{x},\mathcal{D}) = \frac{p(\mathbf{x}|\omega_i,\mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j,\mathcal{D})P(\omega_j|\mathcal{D})}}$$

- We can **separate** the training samples by class into ***c* subsets** $D_1, \ldots, D_c$, samples in D_i belonging to ω_i
- Samples in D_i have **no influence** on $p(\mathbf{x}|\omega_j, D)$ if $i \neq j$.

$$\boxed{P(\omega_i) = P(\omega_i|\mathcal{D}).}$$

$$\boxed{P(\omega_i|\mathbf{x},\mathcal{D}) = \frac{p(\mathbf{x}|\omega_i,\mathcal{D}_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j,\mathcal{D}_j)P(\omega_j)}.}$$

The Parameter Distribution

- Given ***D***, the **prior** distribution can be **updated** to form the **posterior** distribution using the Bayes rule

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

- Where

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

$$\boxed{p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}).}$$

- Although that ***p*(x) is unknown**, **parametric mean** that **p(x|θ) is completely known**

Bayesian Estimation

- The **posterior distribution** p(θ|***D***) can be used to find **estimates for** θ (e.g., the **expected value or mazximum of** of p(θ|***D***) can be used as an estimate for θ).

$$\boxed{\begin{aligned} \widehat{\theta}_{MAP} &= \arg\max_{\theta} \; P(\theta \mid D) \\ &= \arg\max_{\theta} \; P(D \mid \theta) P(\theta) \end{aligned}}$$

- Then, the **conditional density** p(x|D) can be computed as

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}$$

and can be **used in the Bayesian classifier**.

MLEs vs. Bayes Estimates

- Maximum likelihood estimation finds an estimate of θ based on the samples in $\boldsymbol{D}$ but a **different sample** set would **give rise to a different estimate**.
- Bayes estimate takes into account the **sampling variability**.
- We assume that we do not know the true value of θ and instead of taking a **single estimate**, we take a **weighted sum** of the densities $p(x|\theta)$ weighted by the distribution $p(\theta|\boldsymbol{D})$.

The Gaussian Case

- Consider the **univariate** case

$$p(x|\mu) = N(\mu, \sigma^2)$$

where **μ** is the **only unknown parameter** with a **prior distribution**

$$\boxed{p(\mu) = N(\mu_0, \sigma_0^2) \quad (\sigma^2,\ \mu_0 \text{ and } \sigma_0^2 \text{ are all known}).}$$

- This corresponds to **drawing a value for μ** from the population with **density p(μ)**, <u>treating it as the true value</u> in the density p(x|μ), and drawing samples for x from this density.

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

where α is a **normalization factor** that depends on D but is independent of μ.

This equation shows how the observation of a set of training samples **affects our ideas about the true value of μ;**

$$p(\mu|\mathcal{D}) = \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{\boxed{p(x_k|\mu) \sim N(\mu,\sigma^2)}} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{\boxed{p(\mu) \sim N(\mu_0,\sigma_0^2)}}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right],$$

factors that **do not depend on μ** have been absorbed into the constants α, α', and α''.

Given $\mathcal{D} = \{x_1, \ldots, x_n\}$, we obtain

$$p(\mu|\mathcal{D}) \propto \prod_{i=1}^{n} p(x_i|\mu)p(\mu)$$

$$\propto \exp\left[-\frac{1}{2}\left(\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\mu_0}{\sigma_0^2}\right)\mu\right)\right]$$

$$= N(\mu_n, \sigma_n^2)$$

$p(\mu)$ is said to be a ***conjugate prior***

Where

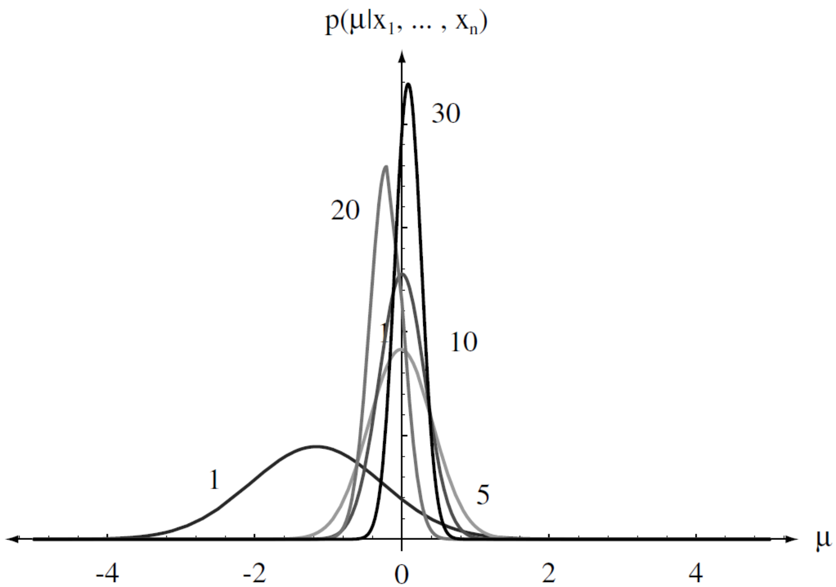
$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

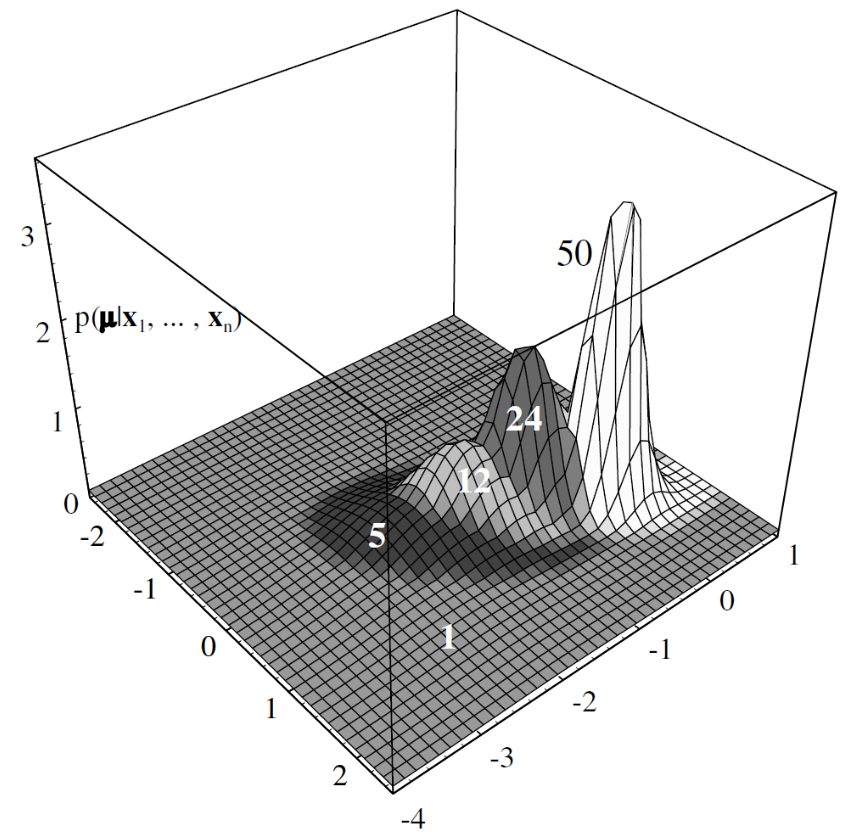
$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}.$$

σ_n^2 uncertainty about nth guess

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

- Since σ_n^2 **decreases monotonically** with n- **approaching σ^2/n** as n approaches infinity
- Each additional observation **decreases our uncertainty** about the true value of μ.
- As n increases, **p(μ|D)** becomes more and more **sharply peaked**, approaching a **Dirac delta** function





The Gaussian Case

- μ_0 is our **best prior guess** and σ_0^2 is the **uncertainty about** this guess.
- μ_n is our best guess **after observing n sample** in **D** and σ_n^2 is the uncertainty about this guess.
- μ_n always lies between $\bar{x}_n$ and μ_0 with coefficients that are non-negative **and sum to one**.

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

  - If $\sigma_0 = 0$, then $\mu_n = \mu_0$ (**no observation can change** our prior opinion).
  - If $\sigma_0 >> \sigma$, then $\mu_n = \bar{x}_n$ (we are **very uncertain** about our prior guess).
  - Otherwise, μ_n approaches $\bar{x}_n$ as n **approaches infinity**

Class-conditional density

- Given the posterior density $p(\mu|\boldsymbol{D})$, the conditional density $p(x|\boldsymbol{D})$ can be computed as

$$
\begin{aligned}
p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})\, d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\
&= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma,\sigma_n),
\end{aligned}
$$

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

- Where

$$\boxed{p(x|\mathcal{D}) = N(\mu_n, \sigma^2+\sigma_n^2)}$$

  - the **conditional mean $\boldsymbol{\mu_n}$** is treated as if it were the true mean,
  - the known **variance is increased** to account for our **lack of exact knowledge** of the mean μ.

The multivariate case

$$p(\mathbf{x}|\boldsymbol{\mu}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where μ is the only unknown parameter with a prior distribution

$p(\boldsymbol{\mu}) = N(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$ ($\boldsymbol{\Sigma}$, $\boldsymbol{\mu_0}$ and $\boldsymbol{\Sigma_0}$ are all known).

Given D = $\{x_1, \ldots, x_n\}$, we obtain

$$p(\boldsymbol{\mu}|\mathcal{D}) \propto \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^T\left(n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma_0}^{-1}\right)\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\mathbf{x}_i + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0}\right)\right)\right].$$

The Multivariate Gaussian Case

It follows that

$$p(\boldsymbol{\mu}|\mathcal{D}) = N(\boldsymbol{\mu_n}, \boldsymbol{\Sigma_n})$$

where

$$\boldsymbol{\mu_n} = \boldsymbol{\Sigma_0}\left(\boldsymbol{\Sigma_0} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\hat{\boldsymbol{\mu}}_{\boldsymbol{n}} + \frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma_0} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu_0},$$

$$\boldsymbol{\Sigma_n} = \frac{1}{n}\boldsymbol{\Sigma_0}\left(\boldsymbol{\Sigma_0} + \frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}.$$

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

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

Class-conditional density

- Given the posterior density p(μ|***D***), the conditional density p(x|***D***) can be computed as

$$p(\mathbf{x}|\mathcal{D}) = N(\boldsymbol{\mu}_{\boldsymbol{n}}, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_{\boldsymbol{n}})$$

- Which can be viewed as the **sum of** a random vector μ with

$$p(\boldsymbol{\mu}|\mathcal{D}) = N(\boldsymbol{\mu}_{\boldsymbol{n}}, \boldsymbol{\Sigma}_{\boldsymbol{n}})$$

- and an **independent random vector y** with

$$p(\mathbf{y}) = N(0, \boldsymbol{\Sigma}).$$

The Bernoulli Case

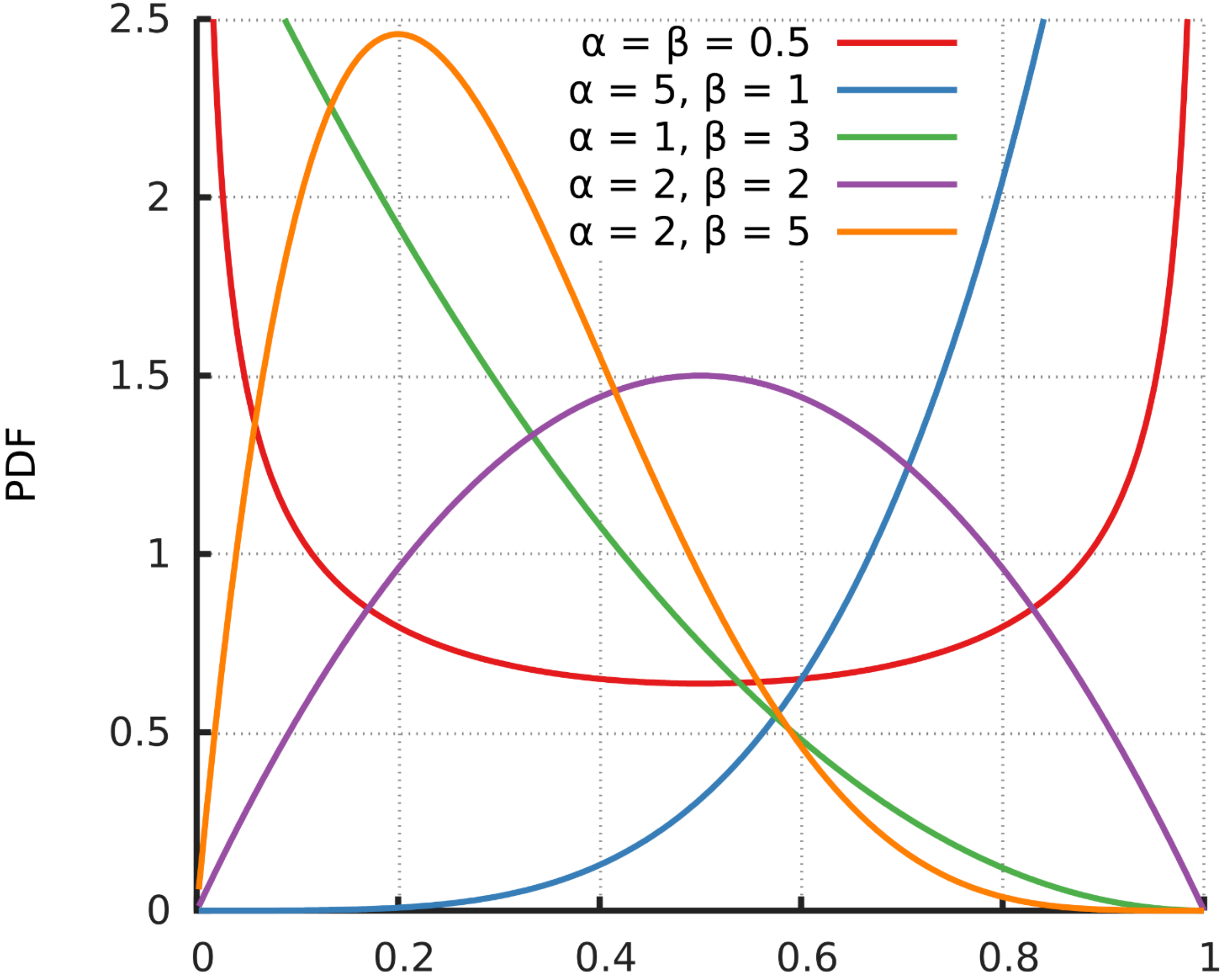
- Consider $P(x|\theta) = \text{Bernoulli}(\theta)$ where θ is the unknown parameter with a prior distribution

  $p(\theta)=\text{Beta}(\alpha,\beta)$ (α and β are both known).

$$f(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

- Given $D = \{x_1, \ldots, x_n\}$, we obtain

$$p(\theta|\mathcal{D}) = \text{Beta}\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i\right).$$



$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

$\alpha > 0$ shape (real)

$\beta > 0$ shape (real)

$x \in [0,1]$ or $x \in (0,1)$

$$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx,$$

$$\mathbf{E}[X] = \frac{\alpha}{\alpha+\beta}$$

Mode	$\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$

$$\operatorname{var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The Bernoulli Case

- The Bayes estimate of θ can be computed as the expected value of p(θ|D), i.e.,

$$
\begin{aligned}
\hat{\theta} &= \frac{\alpha + \sum_{i=1}^{n} x_i}{\alpha + \beta + n} \\
&= \left(\frac{n}{\alpha + \beta + n}\right) \frac{1}{n} \sum_{i=1}^{n} x_i + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) \frac{\alpha}{\alpha + \beta}.
\end{aligned}
$$

Conjugate Priors

- A conjugate prior is one which, when multiplied with the probability of the observation, gives a posterior probability **having the same functional form** as the prior.
- This relationship **allows** the **posterior** to **be used as a prior** in further computations.

pdf generating the sample	*corresponding conjugate prior*
Gaussian	Gaussian
Exponential	Gamma
Poisson	Gamma
Binomial	Beta
Multinomial	Dirichlet

Recursive Bayes Learning

- What about the **convergence** of p(x|D) to p(x)?
- Given $\mathcal{D}^n = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$, for $n > 1$

$$p(\mathcal{D}^n|\boldsymbol{\theta}) = p(\mathbf{x_n}|\boldsymbol{\theta})p(\mathcal{D}^{n-1}|\boldsymbol{\theta})$$

and

$$\boxed{p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\ d\boldsymbol{\theta}},} \qquad p(\boldsymbol{\theta}|\mathcal{D}^n) = \frac{p(\mathbf{x_n}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}^{n-1})}{\int p(\mathbf{x_n}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D}^{n-1})\, d\boldsymbol{\theta}}$$

where

$$p(\boldsymbol{\theta}|\mathcal{D}^0) = p(\boldsymbol{\theta})$$

- Quite useful if the distributions can be represented using **only a few parameters** (**sufficient statistics**).

Recursive Bayes Learning

- Consider the Bernoulli case $P(x|\theta) = \text{Bernoulli}(\theta)$ where $p(\theta)=\text{Beta}(\alpha,\beta)$, the Bayes estimate of θ is

$$\hat{\theta} = \frac{\alpha}{\alpha+\beta}.$$

- Given the training set $D = \{x_1, \ldots, x_n\}$, we obtain

$$p(\theta|\mathcal{D}) = \textbf{Beta}(\alpha+m, \beta+n-m)$$

where

$$m = \sum_{i=1}^{n} x_i = \#\{x_i | x_i = 1, x_i \in \mathcal{D}\}.$$

Recursive Bayes Learning

- The Bayes estimate of θ becomes

$$\hat{\theta} = \frac{\alpha + m}{\alpha + \beta + n}.$$

- Then, given a **new training** set

$$\mathcal{D}' = \{x_1, \ldots, x_{n'}\}$$

- We obtain

$$p(\theta|\mathcal{D}, \mathcal{D}') = \text{Beta}(\alpha + m + m', \beta + n - m + n' - m')$$

- Where

$$m' = \sum_{i=1}^{n'} x_i = \#\{x_i | x_i = 1, x_i \in \mathcal{D}'\}.$$

Recursive Bayes Learning

- The Bayes estimate of θ becomes

$$\hat{\theta} = \frac{\alpha + m + m'}{\alpha + \beta + n + n'}.$$

- Thus, recursive Bayes learning involves **only keeping the counts m** (related to **sufficient statistics** of Beta) and the number of training samples n.

Comparison of MLEs and Bayes estimates

	MLE	*Bayes*
computational complexity	differential calculus, gradient search	multidimensional integration
interpretability	point estimate	weighted average of models
prior information	assume the parametric model $p(\mathbf{x}\|\boldsymbol{\theta})$	assume the models $p(\boldsymbol{\theta})$ and $p(\mathbf{x}\|\boldsymbol{\theta})$ but the resulting distribution $p(\mathbf{x}\|\mathcal{D})$ may not have the same form as $p(\mathbf{x}\|\boldsymbol{\theta})$

If there is **much data** (strongly peaked $p(\theta|D)$) and the prior **$p(\theta)$ is uniform, then the Bayes estimate and MLE** are **equivalent**.

Classification Error

- To apply these results to **multiple classes**, separate the training samples to c subsets $D_1, \ldots, D_c$, with the samples in D_i belonging to class ω_i, and then **estimate each density** $p(x|\omega_i, D_i)$ **separately**.
- Different **sources** of error:
  - **Bayes error**: due to overlapping class-conditional densities (related to the features used; inherent property of the problem and **can never be eliminated**).
  - **Model error**: due to i**ncorrect model**.
  - **Estimation error**: due to estimation from a finite sample (can be reduced by increasing the amount of training data).