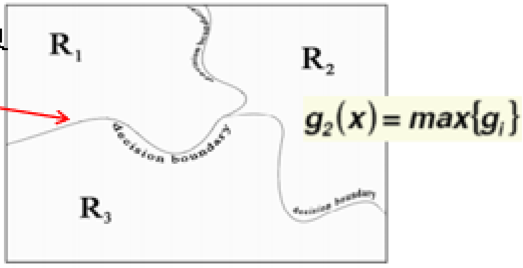
分离器可以表示为一系列的判别函数gi(x)，当对于所有i!=j来说，gi(x)>=gj(x)都成立时，分离器将x分为Ci类

特征空间被分为C个决策域

If $g_i(\boldsymbol{x}) \geq g_j(\boldsymbol{x}),\ \forall j \neq i$ then $\boldsymbol{x}$ is in R_i

Assign $\boldsymbol{x}$ the class C_i

Decision Boundary决策边界



贝叶斯决策规则

先验概率
Prior: A priori probability $p(C_i)$ $\quad \sum_{i=1}^{M} p(C_i) = 1$

证据因子
Evidence: Probability density of feature $\boldsymbol{x}$: $p(\boldsymbol{x})$

似然(类条件概率)
Likelihood: Class-conditional probability density: $p(\boldsymbol{x}|C_i)$

后验概率
Posterior: Probability of class C_i for a given feature value $\boldsymbol{x}$: $p(C_i|\boldsymbol{x})$

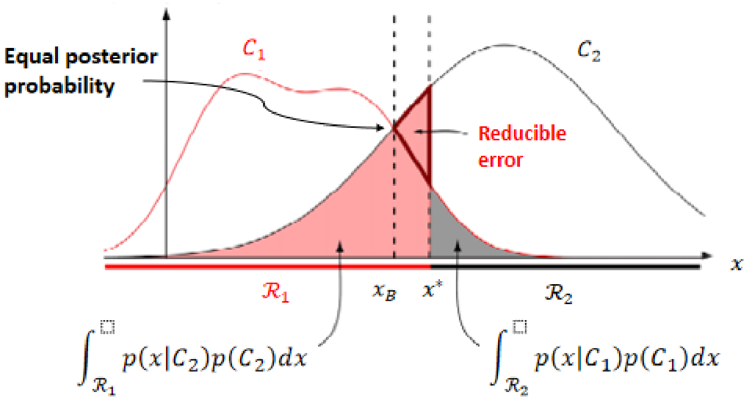
$$p(C_i|\boldsymbol{x}) = \frac{p(C_i)p(\boldsymbol{x}|C_i)}{p(\boldsymbol{x})} \qquad posterior = \frac{prior \times likelihood}{evidence}$$

Bayes Error Rate—Minimum Error Rate

- Bayes decision rule for minimizing the probability of error:

  decide C_1 if $p(C_1|\boldsymbol{x}) > (C_2|\boldsymbol{x})$; otherwise decide C_2

$$p(e) = p(\boldsymbol{x} \in \mathcal{R}_2, C_1) + p(\boldsymbol{x} \in \mathcal{R}_1, C_2)$$



- x^*: nonoptimal decision point.
  - Pink area: the probability of errors for deciding C_1 when the nature is C_2;
  - Gray area: the converse.
- x_B: decision boundary of Bayes decision, where the reducible error is eliminated and the total shaded area is minimum possible (Bayes error rate).

◆ Bayes error rate: the minimum achievable error rate for a classification problem.贝叶斯错误最小化

Why Gaussian高斯分布

■ **Analytical tractability**

- ➢ ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$) are sufficient to uniquely characterize the distribution.（μ，∑）足以唯一地刻画分布。
- ➢ If (Gaussian) x_i's are mutually uncorrelated, then they are independent.如果（高斯）x_i是相互不相关的，那么它们是独立的
- ➢ The marginal and conditional densities are also Gaussian.边际密度和条件密度也是高斯的。

■ **Ubiquity-Frequently observed中心极限定理**

- ➢ Central limit theorem (Many distributions we wish to model are truly close to being normal distributions. 中心极限定理（我们希望建模的许多分布都非常接近正态分布

Covariance Matrix

- The diagonal elements are variances of each feature对角线元素是每个特征的方差
- Relationship between any two features x_i and x_j
  - Independent $\sigma_{ij} = 0$不相关
  - Positive correlation $\sigma_{ij} > 0$正相关
  - Negative correlation $\sigma_{ij} < 0$负相关

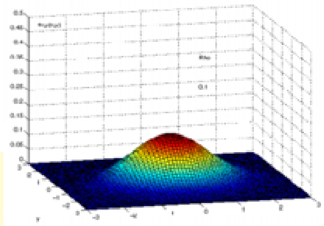
$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_d^2 \end{bmatrix}$$

- If $\boldsymbol{\Sigma}$ is diagonal:对角阵，两两相互独立

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

$$p(\boldsymbol{x}) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} exp\left[-\frac{1}{2}(\frac{x_i-\mu_i}{\sigma_i})^2\right]$$

Mahalanobis Distance马氏距离



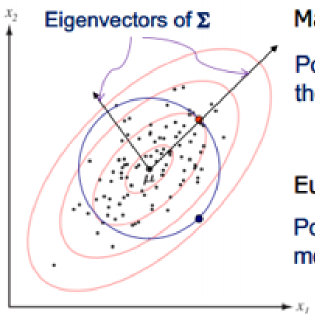
- Probability density function:

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

- Mean vector: $\boldsymbol{\mu}$ Covariance matrix: $\boldsymbol{\Sigma}$
- Mahalanobis distance: $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$

✓ Represents the distance of the test point $\boldsymbol{x}$ from the mean $\boldsymbol{\mu}$.

✓ If $\boldsymbol{\Sigma} = \boldsymbol{I}$, Mahalanobis distance ↔ Euclidean distance.马氏距离等于欧氏距离



Mahalanobis Distance: $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$

Points of equal Mahalanobis distance to the mean lie on an ellipse.

Euclidean Distance: $\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T(\boldsymbol{x}-\boldsymbol{\mu})}$

Points of equal Euclidean distance to the mean lie on a circle.

结论

每个特征的协方差一样就是个线性分离器，反之为二次分离器

- The Bayes classifier for normally distributed classes (general case) is a quadratic classifier.正态分布类的贝叶斯分类器（一般情况）是一个二次分类器。
- The Bayes classifier for normally distributed classes with equal covariance matrices is a linear classifier.具有相等协方差矩阵的正态分布类的贝叶斯分类器是线性分类器。

朴素贝叶斯分离器

朴素贝叶斯分类器是一种基于应用贝叶斯定理（来自贝叶斯统计）和强（朴素）独立性假设的简单概率分类器。

假设类的**特定特征的存在（或不存在）与任何其他特征的存在无关**。

尽管天真的贝叶斯分类器的设计很天真，而且显然过于简化了假设，但它们在许多复杂的现实世界中都能很好地工作。

2004年，对贝叶斯分类问题的分析表明，朴素贝叶斯分类器的有效性明显不合理有一些理论原因

2006年与其他分类方法的综合比较表明，贝叶斯分类在更流行的方法（如增强树或随机森林）中表现更好

优点：

它**只需要少量的训练数据**来估计分类所需的参数（变量的均值和方差）。

因为假设了特征之间相互独立，所以**只需要确定每个特征的方差**，而不需要确定整个协方差矩阵。

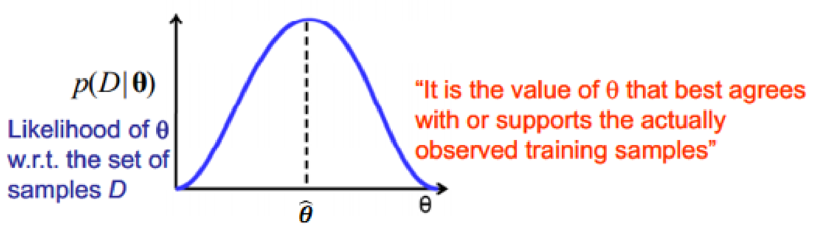
如果有k个类，并且每个p（Fi|C=C）的模型可以用r个参数表示，那么相应的朴素贝叶斯模型具有（k−1）+drk个参数。

Parameter Estimation参数估计

- We can use the *maximum likelihood estimates* of the probabilities.
  - Given a dataset $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, where the n samples are drawn independently from identical distribution $\mathrm{p}(\boldsymbol{x}|\boldsymbol{\theta})$, estimate parameters $\boldsymbol{\theta}$.
  - ML estimate parameters $\boldsymbol{\theta}$ maximizes $\mathrm{p}(\mathcal{D}|\boldsymbol{\theta})$

$\mathcal{D}$ is an i.i.d set

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathrm{p}(\mathcal{D}|\boldsymbol{\theta}) \qquad \mathrm{p}(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} \mathrm{p}(\boldsymbol{x}_k|\boldsymbol{\theta})$$



43

样本矫正

如果给定的类和特征值从未一起出现在训练集中，则**基于频率的概率估计将为零**。

这是有问题的，因为当其他概率相乘时，它会抹去所有信息。

因此，通常希望**在所有概率估计**中**加入小样本校正**，使得没有概率被设置为完全为零

Sample Correction

$$p_\lambda(C=c)=\frac{\sum_{i=1}^{N} I(C=c)+\lambda}{N+K\lambda};$$

K: the total number of classes

$$p_\lambda(F_j=f_j|C=c)=\frac{\sum_{i=1}^{N} I(F_j=f_j,C=c)+\lambda}{\sum_{i=1}^{N} I(C=c)+S_j\lambda};\quad \lambda\geq 0$$

S_j: the total number of possible values of F_j

$\lambda=0$: maximum-likelihood estimation

$\lambda=1$: Laplace Smoothing

Notably, under this correction, we still have, $\sum_{j=1}^{S_j} p_\lambda(F_j=f_j,C=c)=1$

Sj是所有Fj可能取值个数之和

尽管影响深远的独立性假设通常是不准确的，但朴素贝叶斯分类器有几个特性，使其在实践中非常有用

类条件特征分布的解耦意味着每个分布都可以独立地估计为一维分布。这反过来又有助于缓解维度诅咒带来的问题

像MAP决策规则下的所有概率分类器一样，只要正确的类别比任何其他类别都更有可能，它就会达到正确的分类；因此类概率不必被很好地估计。