《模式识别》读书报告 第三部分 分类器与其他工具

第七章 支持向量机

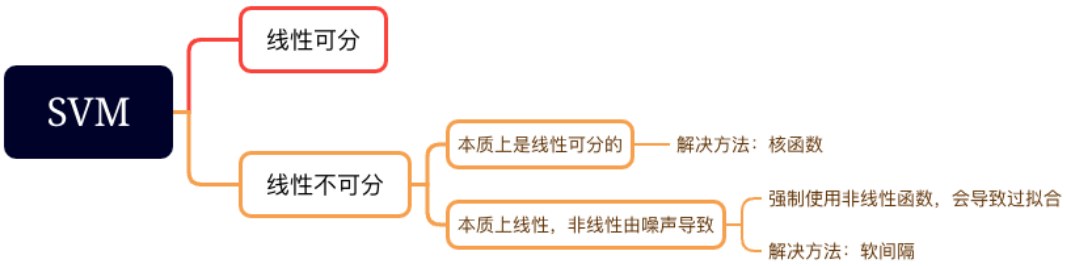
1、SVM 的关键思想

利用训练集来估计概率分布或密度函数，然后有了这些分布贝叶斯决策理论将指导我们如何进行分类。这类概率方法称为生成式方法。

直接估计概率分布或密度函数的方法称为判别式方法。如 SVM。

SVM 不会基于训练数据来对生成式分布或判别式分布进行建模。相反 SVM 希望能直接找到最佳的分类边界，它将 x 的域分为不同的区域。落到同一区域的样本全属于一个类，不同的区域对应不同的类别。

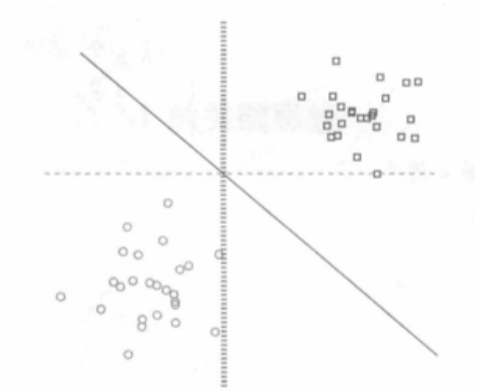
SVM 需要解决的问题是：什么样的边界被认为是好的，或是最好的？



2、以二分类的线性可分问题为例

查找最大间隔的分类器

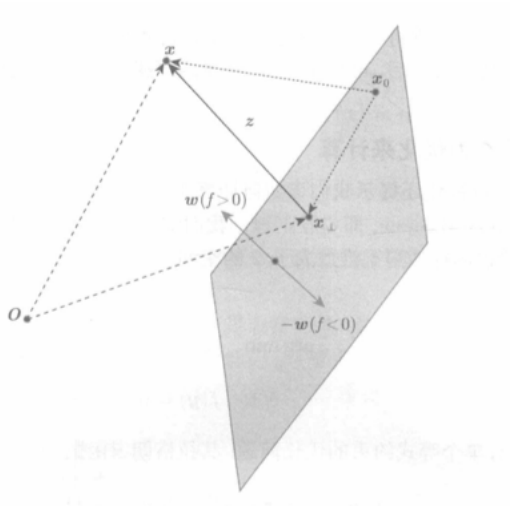
如下所示，实线的那条分类边界明显是优于其他两条的。当出现噪声或者其他扰动时，该分类器往往具有足够的间隔来适应这些变化，分类器的间隔是指从它到最接近它的训练样本之间的距离。



(a) 大间隔边界 vs. 小间隔边界

因为大间隔有利于分类，一些分类器直接最大化间隔，这类分类器被称为最大间隔分类器。一些分类器寻求最大间隔和其他特性的折中，这种称为大间隔分类器。SVM 是一种最大间隔分类器。

3、计算间隔



d 维空间中的一个超平面上的所有点由如公式指定:

$$f(x) = w^T x + b = 0$$

W 指定了超平面的方向。W 的方向$\frac{w}{||w||}$被称为超平面的法向量。

x可被分解为

$$x = x_\perp + z$$

对于超平面上的任一点x_0有

$$z \perp (x_\perp - x_0)$$

$||z||$就是我们要找的距离。

4、最大化间隔

SVM 试图最大化数据集的间隔，即所有训练点的最小间隔。

对于二分类问题令$y = \{-1, +1\}$,要最大化该数据集相对于线性边界$f(x) = w^T x + b$的间隔。

$$\max_{w,\ b} \min_{1 \le i \le n} \frac{f(x_i)}{||w||}$$

$$s.t.\ \ y_i f(x_i) > 0, 1 \le i \le n$$

目标函数还可以被写为:

$$\frac{1}{||w||} \min_{1 \le i \le n} (y_i (w^T x_i + b))$$

原始问题可以转化为下述等价问题:

$$\max_{w,\ b} \frac{\min_{1 \le i \le n} y_i (w^T x + b)}{||w||} = \frac{1}{||w||}$$

$$s.t.\ \ y_i f(x_i) \ge 1, 1 \le i \le n$$

也可以写为:

$$\min_{w,b} \frac{1}{2} w^T w$$

$$s.t.\ \ y_i f(x_i) \ge 1, 1 \le i \le n$$

5、如何求解超平面

引入拉格朗日乘子向量$\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n]$,得到优化问题的对应的拉格朗日函数。各拉

格朗日乘子均为非负数。

$\alpha_i = 0$的观测表示对超平面没有作用，只有$\alpha_i > 0$的观测点才对超平面的系数产生影响，这样的观测点即为支持向量。最大边界超平面完全由支持向量决定。

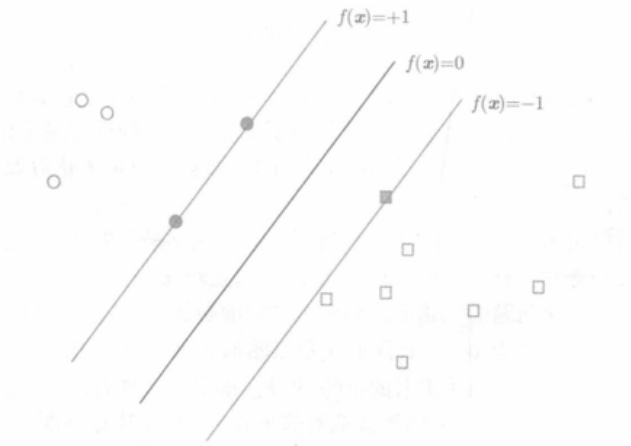


图 7.4　支持向量的示意图（见彩插）

6、求解拉格朗日对偶优化问题

$$\max_{w,b} \min_{1 \le i \le n} L(w, b, \alpha)$$

首先固定α，关于 w 和 b 最小化拉格朗日函数，对 w 和 b 求梯度并置 0，代入原式可得对偶函数：

$$g(\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i y_i w^T x_i + \sum_{i=1}^{n} \alpha_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j$$

对偶优化的具体表示如下：

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$s.t. \quad \alpha_i \ge 0$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

参数 b 可由如下互补松弛条件得到

$$\alpha_i (y_i (w^T x_i + b) - 1) = 0, 1 \le i \le n$$

对于任意支持向量x_i，可得

$$b = y_i - w^T x_i$$

至此，对于待分类样本 x,支持向量机分类器表示为：

$$h(x) = sign[\sum_{i=1}^{n} y_i \alpha_i \, x^T x_i + b]$$

7、线性不可分问题

可以使用一种名为松弛变量的技术，从而扩展线性 SVM 分类器来处理实际问题中的大致线性可分情况。在 SVM 中引入松弛变量ξ_i，作为我们需要为x_i付出的代价，并且与之相关的约束改为：

$$y_i f(x_i) \geq 1 - \xi_i, 1 \leq i \leq n$$
$$\xi_i \geq 0$$

关于 ξ_i 存在三种可能的情况:

- 对于某 x_i，当 $\xi_i = 0$ 成立时，原来的约束仍然成立，不需要付出额外的代价;
- 当 $0 < \xi_i \leqslant 1$ 时，该样本仍被正确分类，但间隔为 $1 - \xi_i < 1$，因此代价为 ξ_i;
- 当 $\xi_i > 1$ 时，该样本被错误分类，其代价仍为 ξ_i.

因此，总代价是

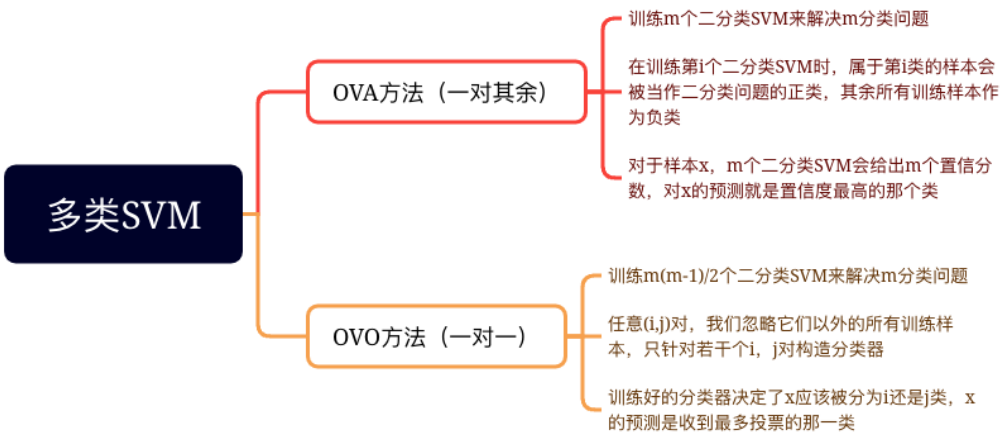
$$\sum_{i=1}^{n} \xi_i .$$

为了让总代价最小，我们将此项作为正则化项添加进目标函数。

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

在不同特性的问题中，我们应该灵活调整不同的 C 以寻求最佳准确率。

8、多类 SVM



m 不大时，上述两种方法都适用，m 很大时 OvA 方法更实用。

9、核 SVM

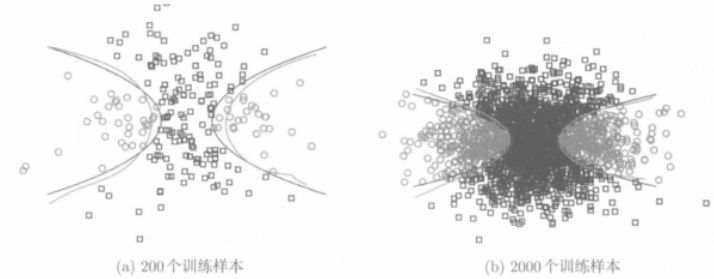


图 7.5　非线性分类器示意图 (见彩插)

如上图线性不可分的情况，核(非线性)SVM 可以很好的逼近真实的决策边界。核方法是非线性学习中非常流行的方法。

SVM 对偶目标中的第二项出现了两个向量的点积，两个向量的点积可被视为相似度的一种度量。我们可以将点积替换为其他的非线性相似度度量方法。

令k表示满足 Mercer 条件的一个非线性函数，则对偶的非线性 SVM 形式化是:

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

10、Mercer 条件

若函数 k(x,y)满足:

1、 若下列不等式对任意平方可积函数 g(x)都成立

$$\iint k(x,y)g(x)g(y)dxdy \geq 0$$

2、 K 函数是对称的，即

$$k(x,y) = k(y,x)$$

那么 k 函数满足 Mercer 条件。

对于一个对称函数 k，和一个样本集 x。当 k 为一个合法的核函数时，它可以将输入 x 从输入空间映射到新的特征空间（通常是更高维度的），从而将低维空间的非线性分类问题转换为高维空间的等效的线性分类问题。

11、常用核函数

线性: $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{y}$,

RBF: $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|^2)$,

多项式: $\kappa(\boldsymbol{x}, \boldsymbol{y}) = (\gamma \boldsymbol{x}^T \boldsymbol{y} + c)^D$.

第八章 概率方法

与概率方法有关的一些模型会对概率函数 p.d.f.或 p.m.f.进行估计，并使用这些概率函数指导我们的决策。

1、贝叶斯定理

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

P(X|Y)被称为似然。它也是一个条件分布，若我们知道 Y=y，那么 X 的分布将变得与其先验 P(X)不同。

贝叶斯定律表明:

$$\text{后验} = \frac{\text{似然} \times \text{先验}}{\text{边缘似然}}$$

2、生成式模型 vs. 判别式模型

如果直接对条件/后验分布 P(Y|X)进行建模，那么这是一个判别式模型。判别式模型不能

采样/生成得到一个服从潜在联合分布的样本对(x,y)。

对联合分布 P(X,Y)进行建模，是生成式模型。通常会对先验分布 P(Y)和类条件分布 P(X|Y)进行建模。

当从联合分布中进行采样的能力不重要时，判别式模型是适用的，并且在实践中通常比生成式模型具有更高的分类精度。

3、参数化方法

如果我们假设分布为正态分布，那么它的 p.d.f.就具有固定形式：

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

该函数由两个参数指定，均值μ和协方差矩阵Σ。因此要估计该分布就是估计其参数。

给定一个数据集 D={x1,x2,⋯,xn},若服从正态分布，那么它的极大似然估计（ML 估计）为

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$$

$$\Sigma_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu}_{\mathrm{ML}})(\boldsymbol{x}_i-\boldsymbol{\mu}_{\mathrm{ML}})^T.$$

ML 估计（极大似然估计）和 MAP 估计（最大后验估计）都是参数化方法。因为我们会假设特定的函数形式（如正态分布的 p.d.f.）并且只对这些函数中的参数进行估计。

当一个连续分布的函数形式未知时，我们可以使用 GMM（高斯混合模型）代替。

$$p(x) = \sum_{i=1}^{K}\alpha_i N(x;\mu_i,\Sigma_i)$$

α_i非负且满足$\sum_{i=1}^{K}\alpha_i = 1$，$N(x;\mu_i,\Sigma_i)$为第 i 个多元高斯分量。该 GMM 分布为一个合法连续分布。

4、非参数化方法

非参数化方法没有对密度函数假设一个特定的函数形式，非参数化方法用训练样本来估计定义域中任意点处的密度。“非参数化”意味着没有假设参数化的函数形式，而并不是说不需要参数。

当训练样本的数量增加时，非参数化模型中参数的数量通常也会增加，非参数化方法通常面临高额的计算代价。

5、参数化估计

5.1 最大似然估计

我们通常会定义一个似然函数：

$$\ell(\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

若概率密度函数中包含指数函数，我们对其求对数，对数似然函数定义如下：

$$\ell\ell(\theta) = \ln\big(\ell(\theta)\big) = \sum_{i=1}^{n}\ln\ p(x_i|\theta)$$

最大似然估计求解以下的优化问题

$$\theta_{ML} = \arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \ell\ell(\theta)$$

以正态分布为例，我们可以对其求偏导并置 0，就可以得到：

$$\mu_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} x_i\,,$$
$$\sigma^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\mathrm{ML}})^2\,.$$

5.2 最大后验估计

如果我们有足够多的样本，ML 估计可以很准确，但是如果只有少量样本时，ML 估计通常会遭遇不准确的结果。一种解决方法是引入我们关于参数的领域知识。

若我们知道均值μ在 5.5 附近，那么这个知识就可以被翻译为一个先验分布：

$$p(\theta) = p(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left(-\frac{(\mu-5.5)^2}{2\sigma_0^2}\right)$$

其中σ_0是一个比较大的数字，在这个例子中我们假设不存在关于σ的先验知识，但先验的假设μ服从高斯分布，其均值为 5.5，并且方差很大。那么最大后验估计求解如下的问题

$$\arg\max_{\theta} p(\theta)\ell(\theta) = \arg\max_{\theta}(\ln(p(\theta)) + \ell\ell(\theta))$$

优化过程与 ML 估计类似。MAP 将先验知识和训练数据同时纳入考虑，当训练数据很小时，先验$\ln(p(\theta))$可能会起到很重要的作用。样本数量很大，那么$\ln(p(\theta))$的作用就会被$\ell\ell(\theta)$稀释。

5.3 贝叶斯参数估计

在贝叶斯观点和贝叶斯参数估计中，θ是一个随机变量，这意味着它的最优估计不是一个固定值，而是一个完整的分布。贝叶斯估计的输出是$p(\theta|D)$。

贝叶斯定理是贝叶斯估计的关键。

给定一个数据集$D = \{x_1, x_2, \ldots, x_n\}$，有 n 个随机变量$X_i$,他们是独立同分布的，并且$x_i$是从$X_i$中抽样得到的。因此 D 为某随机变量数组的一个样本。假设这些随机变量均服从相同的正态分布$N(\mu,\sigma^2)$，假设σ已知，我们只需要估计μ，因此$\theta = \mu$。

因为$\theta(\mu)$是一个随机变量，我们假设其先验分布为

$$p(\mu) = N(\mu;\mu_0,\sigma_0^2)$$

假设μ_0和σ_0^2均已知，我们需要估计$p(\mu|D)$，由贝叶斯定理：

$$\begin{aligned} p(\mu|D) &= \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)\,\mathrm{d}\mu} \\ &= \alpha p(D|\mu)p(\mu) \\ &= \alpha\prod_{i=1}^{n} p(x_i|\mu)p(\mu)\,, \end{aligned}$$

其中$\alpha = \frac{1}{\int p(D|\mu)p(\mu)d\mu}$为一个规范化常数。

根据正态分布的性质，我们有

$$p(\mu|D) = N(\mu_n, \sigma_n^2)$$

其中

$$(\sigma_n^2)^{-1} = (\sigma_0^2)^{-1} + \left(\frac{\sigma^2}{n}\right)^{-1},$$
$$\mu_n = \frac{\sigma^2/n}{\sigma_0^2 + \sigma^2/n}\mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2/n}\mu_{\mathrm{ML}}.$$

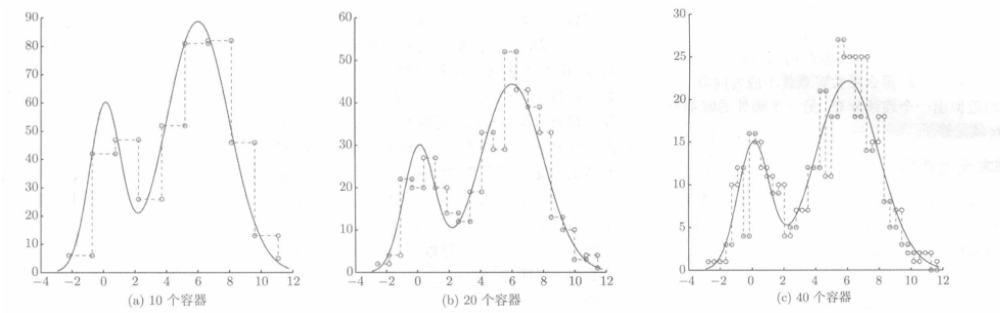
贝叶斯估计说明：当只有很少的训练样本时，一个适当的先验分布是有帮助的，当有足够样本时，先验分布可以被安全的忽略。

6、非参数化估计

非参数化估计不会对密度函数形式做任何假设。

6.1 直方图近似

我们利用直方图来近似原始的 p.d.f.曲线。



但是直方图近似面临着以下的问题

1、 没有连续的估计

2、 维数灾难

3、 需要找到合适的容器宽度

6.2 KDE 核密度估计

令 K 表示一个非负核函数，其积分为 1($\int K(x)dx = 1$)。此外，我们还要求$\int xK(x)dx = 0$那么，核密度估计为：

$$p_{KDE}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}K(\frac{(x - x_i)}{h})$$

注意，

这里所说的核方法与 SVM 中的核有不同含义，尽管有些函数在两种方法中都是有效的核函数（如 RBF 核/高斯核）

参数 h>0 与直方图估计中的容器宽度有着类似作用。在 KDE 中，该参数被称为带宽。

可以推导出$p_{KDE}(x) \geq 0$，以及$\int p_{KDE}(x)dx = 1$,因此核密度估计是一个有效的 p.d.f.

6.3 常用核函数

1、 Epanechnikov 核

在最小平方误差意义上，Epanechnikov 核被证明是最优的核函数

$$K(x)=\begin{cases}\frac{3}{4\sqrt{5}}(1-\frac{x^2}{5}), if|x|<\sqrt{5}\\ 0 \quad ,otherwise\end{cases}$$

该核函数有一个有限的支撑。

2、 高斯核

实践中高斯核更受欢迎，它有无限的支撑

$$K(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right),-\infty<x<+\infty$$

带宽为 h 时，KDE 为

$$p_{KDE}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2}\pi h}\exp\left(-\frac{(x-x_i)^2}{2h^2}\right)$$

即使高斯核不是最优的，如果仔细选择带宽 h，高斯核与 Epanechnikov 核之间的误差差异会很小。

6.4 带宽的选择

在对需要估计的密度函数 p 与核 K 做一些相当弱的限制条件后，理论上的最优带宽为：

$$h^{\star}=\frac{c_1^{-2/5}c_2^{1/5}c_3^{-1/5}}{n^{1/5}}$$

其中$c_1=\int x^2K(x)\mathrm{d}x, c_2=\int K^2(x)\mathrm{d}x, c_3=\int(p''(x))^2\mathrm{d}x$。

c_3d 值很难被可靠的估计，但是如果 p(x)是一个正态分布，那么存在一个实践中可用的规则

$$h^{\star}\approx\left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5}\approx 1.06\hat{\sigma}n^{-1/5}$$

其中，$\hat{\sigma}$是在训练集上估计的标准差。

然而，当数据与高斯不相似时，上述公式可能会导致非常差的密度估计结果。在这种情况下，交叉验证可以被用于密度估计。

6.5 多变量 KDE

H 为带宽矩阵，H 是对称正定的，K 是核函数，它是中心化且对称的。我们期望$x=x_i$时$K(x-x_i)$是最大的。如果我们使用多变量高斯核，则有

$$p_{\mathrm{KDE}}(\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{d/2}|H|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_i)^TH^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)\right)$$

所有的高斯分量共享相同的协方差矩阵(带宽矩阵 H)。在实践中通常会假设一个对角带宽矩阵$H=\mathrm{diag}\,(d_1,d_2,\ldots,d_n)$。

对角 GMM 是连续分布的通用逼近器。因此，我们希望对角矩阵 H 也能得到潜在核密度函数 p(x)的准确的近似。对角多变量 KDE，以高斯核为例

$$p_{\mathrm{KDE}}(\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{d/2}\prod_{j=1}^{d}h_j}\prod_{j=1}^{d}\exp\left(-\frac{\left(x_j-x_{i,j}\right)^2}{2h_j^2}\right)$$

其中 x_j 是新样本 $\boldsymbol{x}$ 的第 j 个维度, $x_{i,j}$ 是第 i 个训练样本 $\boldsymbol{x}_i$ 的第 j 维。

第九章 距离度量与距离变换

1、距离度量

最常用的距离度量是欧式距离

$$d(\boldsymbol{x},\boldsymbol{y}) = \| \boldsymbol{x}-\boldsymbol{y} \| = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$$

对于任意$\boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \mathcal{X}$，欧式距离满足：

i) $d(\boldsymbol{x},\boldsymbol{y}) \geqslant 0$ (非负性);
ii) $d(\boldsymbol{x},\boldsymbol{y}) = d(\boldsymbol{y},\boldsymbol{x})$ (对称性);
iii) $d(\boldsymbol{x},\boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$ (同一性);
iv) $d(\boldsymbol{x},\boldsymbol{z}) \leqslant d(\boldsymbol{x},\boldsymbol{y}) + d(\boldsymbol{y},\boldsymbol{z})$ (三角不等式).

任意满足上述四个条件的映射$f: \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$被称为度量。除了欧式距离外，还有其他的距离度量方式如离散度量和ℓ_p度量。

离散度量在比较两个类别值时很有用：

$$p(x,y) = \begin{cases} 1 & if\ x \neq y \\ 0 & if\ x = y \end{cases}$$

2、向量范数和度量

函数 f 是一个向量范数的条件是它要满足下列三个性质：

i) $f(c\boldsymbol{x}) = |c| f(\boldsymbol{x})$ (齐次性);
ii) $f(\boldsymbol{x}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{0}$ (非负性);
iii) $f(\boldsymbol{x}+\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + f(\boldsymbol{y})$ (三角不等式).

向量范数总是非负的。

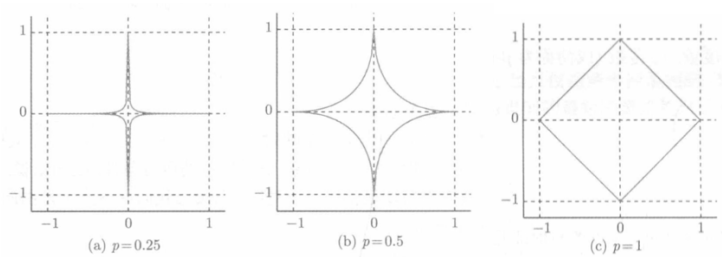
3、ℓ_p范数和ℓ_p度量

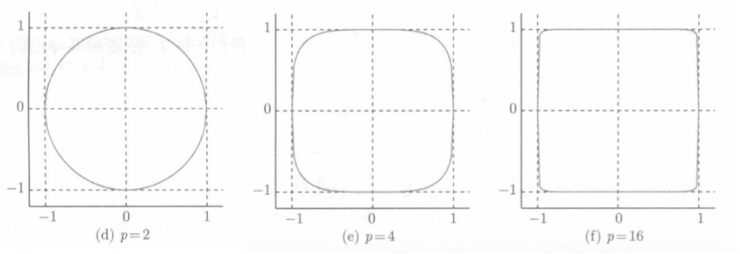
ℓ_p范数（p-范数）在 p>=1 时在$\mathbb{R}^d$空间的定义为：

$$||x||_p = \left(\sum_{i=1}^{d}|x_i|^p\right)^{\frac{1}{p}}$$

当 p 为无理数且 x 为负数时，x^p无定义，因此绝对值符号不可省略。

下图为不同 p 值下满足$||x||_p = 1$d 轮廓





ℓ_∞ 距离衡量了任意一维的最大距离：

$$d_\infty(\boldsymbol{x},\boldsymbol{y}) = \max\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_d - y_d|\}.$$

当 $p > q > 0$, 我们有

$$\|\boldsymbol{x}\|_p \leqslant \|\boldsymbol{x}\|_q .$$

也就是说，当 p 增加时，ℓ_p范数将减小或者不变。

4、距离度量学习

平方距离

$$d_A^2(x,y) = (x-y)^T A(x-y)$$

平方马氏距离是上式的一个特例，其中 A 固定为Σ^{-1},当数据不服从高斯分布时，使用协方差矩阵的逆不是最优的。因此，距离度量学习试图从训练数据中学习到一个更好的矩阵 A，在这种情况下我们用$||x-y||_A$来表示 x 和 y 之间学到的距离。

$$||x-y||_A^2 = (x-y)^T A(x-y)$$

在二分类问题中，如果给定一个训练集 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$，我们可以通过优化如下问题学习到一个合适的距离度量：

$$\min_{A\in\mathbb{R}^d\times\mathbb{R}^d} \sum_{\substack{1\leqslant i<j\leqslant n \\ y_i=y_j}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A^2 \tag{9.16}$$

$$\text{s.t.} \sum_{\substack{1\leqslant i<j\leqslant n \\ y_i\neq y_j}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_A^2 \geqslant 1 \tag{9.17}$$

$$A \succeq 0 . \tag{9.18}$$

其中 A 是半正定的。

5、均值作为相似度度量

广义平均

给定两个分布，下图中位置 v 处的相似度贡献为

$$\min\left(p_X(v), p_y(v)\right)$$

相似度为

$$\int \min\Big(p_X(v), p_y(v)\Big)\, dv$$

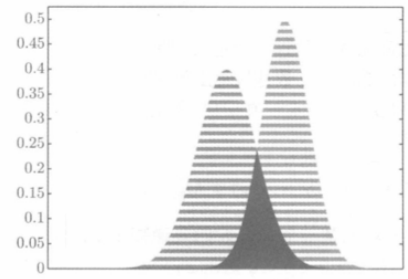


图 9.3　两个分布的相似度图示 (见彩插)

幂平均

一组正值$x_1, x_2, \ldots, x_n$，指数为 p 的广义平均，也被称为幂平均，定义为

$$M_p(x_1, x_2, \ldots, x_n) = (\frac{1}{n}\sum_{i=1}^{n} x_i^p)^{1/p}$$

6、幂平均核

给定两个非负向量 x 和 y，其中向量的元素都是非负的，幂平均核定义为

$$M_p(x, y) = \sum_{i=1}^{d} M_p(x_i, y_i)$$

我们想比较两个分布，如两个表示为直方图的分布时，每个分布值非负，幂平均核家族通常会比常用核（如内积、PBF 或多项式核）更好的相似度度量。

$$M_0(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} \sqrt{x_i y_i}\,, \qquad \text{(Hellinger 核)}$$

$$M_{-1}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} \frac{2x_i y_i}{x_i + y_i}\,, \qquad (\chi^2 \text{ 核})$$

$$M_{-\infty}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} \min(x_i, y_i)\,. \qquad \text{(直方图相交核)}$$

7、线性回归

线性回归基于几个简单的假设：首先，假设自变量 x 和因变量 y 之间的关系是线性的，即 y 可以表示为 x 中元素的加权和，这里通常允许包含观测值的一些噪声； 其次，我们假设任何噪声都比较正常，如噪声遵循正态分布。

$$y_i = x_i^T \beta + \epsilon_i$$

线性回归的优化问题如下

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 = \arg\min_{\boldsymbol{\beta}} \left(\boldsymbol{\beta}^T X^T X \boldsymbol{\beta} - 2\boldsymbol{y}^T X \boldsymbol{\beta}\right).$$

线性回归的解为：

$$\boldsymbol{\beta}^* = (X^T X)^{-1} X^T \boldsymbol{y}\,.$$

8、特征规范化

如果不同尺度间的数值大小差异没有数据生成过程中的特定性质的支撑，一些特征维度

错误的尺度会在机器学习和模式识别中造成严重后果。

我们可以通过如下的手段将第 j 维的数值范围规范化到[0,1]

$$\hat{x}_{i,j} = \frac{x_{i,j} - x_{min,j}}{x_{max,j} - x_{min,j}}$$

规范化到[-1,1]:

$$\hat{x}_{i,j} = 2(\frac{x_{i,j} - x_{min,j}}{x_{max,j} - x_{min,j}} - 0.5)$$

如果存在某个维度的值全部相同，表示该维度没有任何用处直接丢弃。

如果一个向量有很多维都是 0，那么就被称为是稀疏的，下式将 0 规范化到

$$-\frac{x_{min,j}}{x_{max,j} - x_{min,j}}$$

若我们明确知道某个维度服从高斯分布，那么我们将其规范化到标准高斯:

$$\hat{x}_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}$$

9、Softmax 变换

softmax 函数能够将未规范化的预测变换为非负数并且总和为 1，同时让模型保持 可导的性质。 为了完成这一目标，我们首先对每个未规范化的预测求幂，这样可以确保输出非负。为了确保最终输出的概率值总和为 1，我们再让每个求幂后的结果除以它们的总和。如下式:

$$\hat{\mathbf{y}} = \mathrm{softmax}(\mathbf{o})其中\hat{y}_j = \frac{\exp(o_j)}{\sum_k \exp(o_k)}$$

这里，对于所有的j总有$0 \leq \hat{y}_j \leq 1$。 因此，$\hat{\mathbf{y}}$可以视为一个正确的概率分布。 softmax 运算不会改变未规范化的预测$\mathbf{o}$之间的大小次序，只会确定分配给每个类别的概率。 因此，在预测过程中，我们仍然可以用下式来选择最有可能的类别。

$$\underset{j}{\mathrm{argmax}}\hat{y}_j = \underset{j}{\mathrm{argmax}}o_j.$$

尽管 softmax 是一个非线性函数，但 softmax 回归的输出仍然由输入特征的仿射变换决定。 因此，softmax 回归是一个线性模型（linear model）。