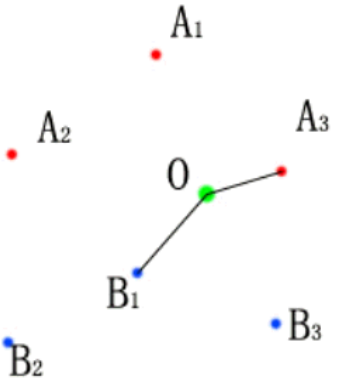
3.4.2.1 最近邻法错误率分析

其实近邻法的错误率是比较难算的，因为训练样本集的数量总是有限的，有时多一个少一个训练样本对测试样本分类的结果影响很大。譬如图中



红点表示A类训练样本，蓝点表示B类训练样本，而绿点O表示待测样本。假设以欧氏距离来衡量，O的最近邻是A_2，其次是B_1，因此O应该属于A类，但若A_2被拿开，O就会被判为B类。这说明计算最近邻法的错误率会有偶然性，也就是指与具体的训练样本集有关。同时还可看到，计算错误率的偶然性会因训练样本数量的增大而减小。因此人们就利用训练样本数量增至极大，来对其性能进行评价。这要使用渐近概念，以下都是在渐近概念下来分析错误率的。

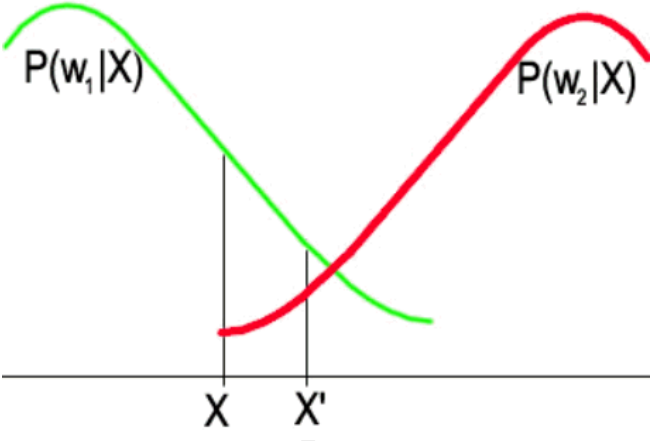


图 3.17

当最近邻法所使用的训练样本数量N不是很大时，其错误率是带有偶然性的。为了说明这一点我们拿图3.17所示一个在一维特征空间的两类别情况来讨论。图中X表示一特测试样本，而X'是所用训练样本集中X的最邻近者，则错误是由X与X'分属不同的类别所引起的。由于X'与所用训练样本集有关，因此错误率有较大偶然性。但是如果所用训练样本集的样本数量N极大，即$N\to\infty$时，可以想像X'将趋向于X，或者说处于以X为中心的极小邻域内，此时分析错误率问题就简化为在X样本条件下X与一个X(X'的极限

条件)分属不同类别的问题。如果样本X的两类别后验概率分别为$P(\omega_1|X)$与$P(\omega_2|X)$，那么对X值，在$N\to\infty$条件下，发生错误决策的概率为：

$$\lim_{n\to\infty} P_N(e|X) = 1-\sum_{i=1}^{c} P^2(\omega_i|X) \tag{3-64}$$

当训练样本数量无限增多时，一个测试样本X的最近邻在极限意义上讲就是X本身。如果在X处对某一类的的后验概率为$P(\omega_1|X)$，则另一类为$1- P(\omega_1|X)$。那么当前测试样本与它的最近邻都属于同一类才能分类正确，故正确分类率为 $\sum_{i=1}^{c} P^2(\omega_i|X)$ ，故有(3-64)式。

而在这条件下的平均错误率

$$P = \lim_{N\to\infty} P_N(e) = \lim_{N\to\infty} \int P_N(e|X)p(X)dX$$

$$= \int \lim_{N\to\infty} P_N(e|X)p(X)dX = \int [1-\sum_{i=1}^{c} P^2(\omega_i|X)]p(X)dX \tag{3-65}$$

P称为渐近平均错误率，是$P_N(e)$在$N\to\infty$的极限。

为了与基于最小错误率的贝叶斯决策方法对比，下面写出贝叶斯错误率的计算式。

基于最小错误率贝叶斯决策的错误率是出错最低限，因此要与它作比较。

$$P^* = \int P^*(e|X)p(X)dx \tag{3-66}$$

其中

$$P^*(e|X) = 1-P(\omega_m|X) \tag{3-67}$$

而

$$P^*(\omega_m|X) = \max_i [P(w_i|X)] \tag{3-68}$$

如果用图3.17中的例子，则从(3-67)可得

$$P^*(e|X) = 1-P(\omega_1|X) \tag{3-69}$$

而从(3-64)得

$$\lim_{N\to\infty} P_N(e|X) = 1-P^2(\omega_1|X)-P^2(\omega_2|X) \tag{3-70}$$

如果用(3-70)减去(3-69)，并写成△P，则有

$$\triangle P = P(\omega_1|X)[1-P(\omega_1|X)] - P^2(\omega_2|X)$$

$$= P(\omega_2|X)[P(\omega_1|X)-P(\omega_2|X)] \tag{3-71}$$

从(3-71)式可见在一般情况下△P是大于零的值，只要$P(\omega_1|X)>P(\omega_2|X)>0$。有以下两种例外情况△P＝0，这两种情况是$P(\omega_1|X)=1$的情况或$P(\omega_1|X)=P(\omega_2|X)=1/2$。

请想一下，什么情况下$P(\omega_1|X)=1$或$P(\omega_2|X)=1$? $P(\omega_1|X)= P(\omega_2|X)$会出现什么什么情况?

答：一般来说，在某一类样本分布密集区，某一类的后验概率接近或等于1。此时，基于最小错误率贝叶斯决策基本没错，而近邻法出错可能也很小。而后验概率近似相等一般出现在两类分布的交界处，此时分类没有依据，因此基于最小错误率的贝叶斯决策也无能为力了，近邻法也就与贝叶斯决策平起平坐了。

从以上讨论可以看出，当$N\to\infty$时，最近邻法的渐近平均错误率的下界是贝叶斯错误率，这发生在样本对某类别后验概率处处为1的情况或各类后验概率相等的情况。

在其它条件下，最近邻法的错误率要高于贝叶斯错误率，可以证明以下关系式成立

$$P^* \le P \le P^*(2-\frac{C}{C-1}P^*) \tag{3-72}$$

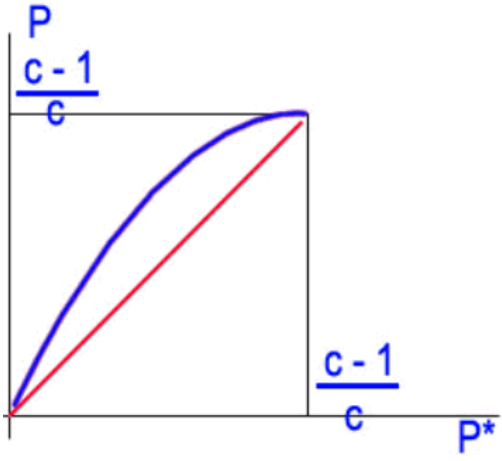


图 3.18

即最近邻法的渐近平均错误率的上下界分别为贝叶斯错误率 P^* 及 $P^*(2-\frac{C}{C-1}P^*)$。图3.18表示了这种关系。由于一般情况下 P^* 很小，因此(3-72)又可粗略表示成

$$P^* \le P \le 2P^*$$

因此可以说最近邻法的渐近平均错误率在贝叶斯错误率的两倍之内。从这点说最近邻法是优良的，因此它是模式识别重要方法之一。

3.4.2.2 k-近邻法错误率分析

这一节不作基本要求。

以上我们从定性分析的角度讨论了最近邻法错误率问题，下面以同样的方法更简略地讨论k-近邻法的渐近平均错误率。对于两类别问题，式(3-64)可以改写成

$$\lim_{N\to\infty} P_N(e \mid X) = P(\omega_1 \mid X)P(\omega_2 \mid X) + P(\omega_2 \mid X)P(\omega_1 \mid X) \tag{3-73}$$

推广到k-邻域的情况，则错误出现在k个邻域样本中，正确的类别所占样本未过半数，得到

$$P^k_{N\to\infty}(e \mid X) = P(\omega_1 \mid X)\sum_{j=0}^{(k-1)/2} C_k^j P(\omega_1 \mid X)P(\omega_2 \mid X)^{k-j} + P(\omega_2 \mid X)\sum_{j=0}^{(k-1)/2} C_k^j P(\omega_2 \mid X)P(\omega_1 \mid X)^{k-j} \tag{3-74}$$

$$C_k^j = \frac{k!}{j!(k-j)!}$$

其中

k邻域出错是指某类样本的k近邻中同类训练样本占少数，仅占一个两个，至多(k-1)/2个，因此这些情况都要考虑，计算就相当复杂了。

将(3-74)与(3-73)相比较，(3-73)相当于(3-74)中k＝1的情况，而在(3-74)中当k增大时 $P^k_{N\to\infty}(e \mid X)$ 是单调递减的。因此可以得出结论，在N→∞的条件下，k-近邻法的错误率要低于最近邻法，图3-19图示了不同k值时的错误率情况。

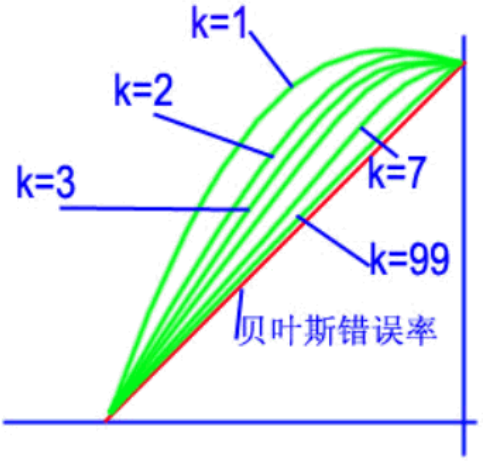


图 3.18

从图中也可看出，无论是最近邻法，还是k-近邻法，其错误率的上下界都是在一倍到两倍贝叶斯决策方法的错误率范围内。

链接： http://202.197.191.206:8080/30/text/chapter03/3_4t2.htm