第七讲 基于近邻的分类方法

牟克典

2021年4月28日

在日常生活中我们也经常采用“近朱者赤、近墨者黑”的方式，利用事物之间的相似性来对新事物进行相关预测.

本将所要介绍的基于近邻的分类方法就是这样一类方法.

- 依据训练样本集中与新数据实例比较相似的若干样本（一般称之为新数据的近邻）的类标记来预测新数据的类标记.
- 这类方法通常并不显式地从训练数据出发探索数据的特征向量和相应的类标记之间的明确联系.
- 没有显式的模型学习或训练过程，一般被称为“惰性”学习方法.

基于近邻的分类方法的核心问题:

- 如何度量数据之间的相似性.
- 选择哪些和新数据实例相似的样本.
- 如何利用选定样本的类标记来预测新实例的类标记.

给定数据的相似性度量

- 如果选择与新数据最相近的k 个样本点来对新数据的类别进行预测，这就是典型的k- 近邻法.
- 特别地，如果我们直接将新数据的类标记预测为与其最相近的训练样本的类别，这就是最近邻法.

概要

k-近邻法的主要思想

- 对于一个新的数据x的类别预测来说，
  - 先从给定的训练样本集中找出和x 最相近的k个样本，
  - 然后通过对这k个近邻样本的类别标记进行多数占优的投票方式来确定x 的类标记.
- 对给定k值的k-近邻法来说，核心问题就是如何从训练样本集中确定与x最相近的k个样本.

特征空间是n维实向量空间$\mathbf{R}^n$的情形下，通常使用两个特征向量的Minkowski距离来度量两个特征向量的相似性.

设$x_i, x_j \in \mathcal{X} = \mathbf{R}^n$, 则$x_i$和$x_j$之间的Minkowski 距离$dist_p(x_i, x_j)$定义为

$$dist_p(x_i, x_j) = \|x_i - x_j\|_p = \left(\sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|^p \right)^{\frac{1}{p}},$$

这里$p \geq 1$, $x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^T$ 且 $x_j = (x_j^{(1)}, x_j^{(2)}, \cdots, x_j^{(n)})^T$.

- 当$p=1$时，$dist_1(x_i,x_j)$被称为曼哈顿（Manhattan）距离或者街区距离，即

$$dist_1(x_i,x_j)=\|x_i-x_j\|_1=\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|$$

- 当$p=2$时，$dist_2(x_i,x_j)$就是欧氏距离，即

$$dist_2(x_i,x_j)=\|x_i-x_j\|_2=\sqrt{\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^2}.$$

- 当$p=\infty$时，$dist_\infty(x_i,x_j)$就是切比雪夫（Chebyshev）距离，即

$$dist_\infty(x_i,x_j)=\|x_i-x_j\|_\infty=\max_{1\le l\le n}|x_i^{(l)}-x_j^{(l)}|.$$

二维空间$\mathbf{R}^2$中两个数据点x_1和x_2之间的曼哈顿距离、欧氏距离和切比雪夫距离的比较示意:

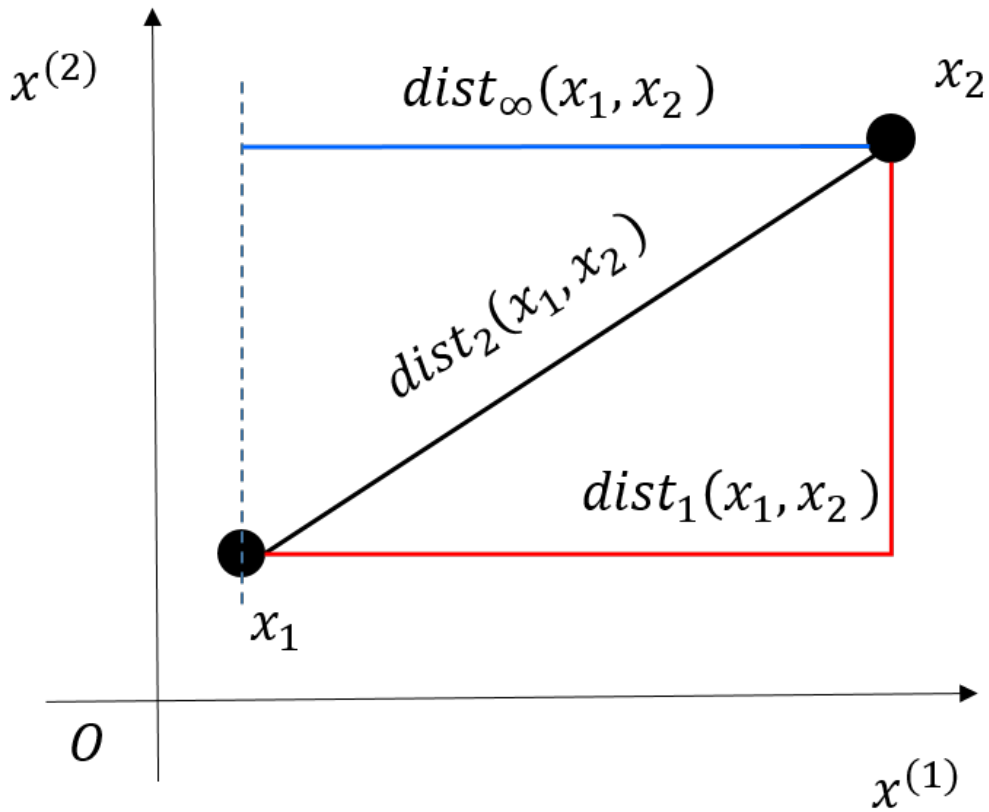

Figure: 曼哈顿距离、欧氏距离和切比雪夫距离

$\mathbf{R}^2$空间中在不同度量下所有到x_1的距离等于$dist_p(x_2, x_1)$的数据点的分布情况:

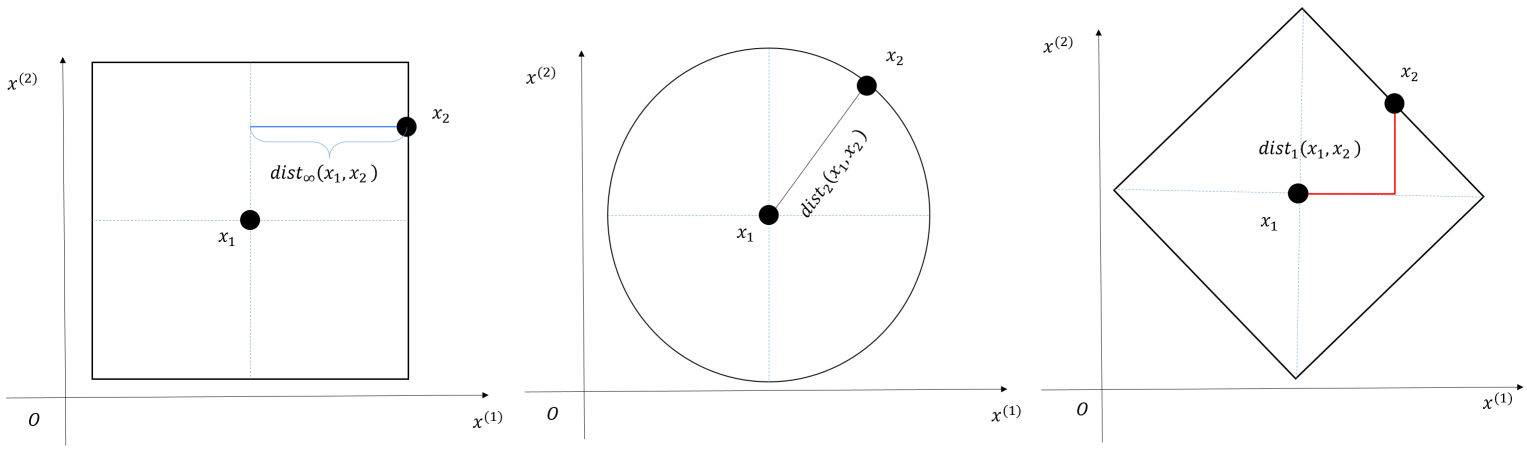

Figure: 满足$dist(x, x_1) = dist(x_2, x_1)$的点$x$的分布情况

如果数据点的各维特征的重要性不同，可以使用加权Minkowski距离$dist_{wp}$来度量两个特征向量的相似性：

$$dist_{wp}(x_i, x_j) = \left(\sum_{l=1}^{n} w_l |x_i^{(l)} - x_j^{(l)}|^p \right)^{\frac{1}{p}},$$

这里$w_l \geq 0$用于表示第l维特征的重要程度，通常我们要求

$$\sum_{l=1}^{n} w_l = 1.$$

给定

- 训练样本集$D = \{(x_i, y_i)\}_{i=1}^{N}$, 其中$x_i \in \mathcal{X} = \mathbf{R}^n$, $y_i \in \mathcal{Y} = \{c_m\}_{m=1}^{M}$, $i = 1, 2, \cdots, N$.
- 给定距离度量为$dist$.

k-近邻法对新实例点x所属类别y的预测由下列两步组成:

(1) 基于距离度量$dist$, 找出训练样本集D中与x最邻近的k个点构成的x的邻域$N_k^{dist}(x)$;

(2) 对$N_k^{dist}(x)$中的样本点采用如下多数占优的投票规则决定x所属的类y:

$$y = \operatorname*{argmax}_{c_m} \sum_{x_i \in N_k^{dist}(x)} I(y_i = c_m).$$

距离度量对预测结果的影响

给定$\mathcal{Y}=\{+1,-1\}$的二类分类问题

- 正类样本集 $D_{+1}=\{x_1=(2,2)^T, x_2=(0,4)^T\}$,
- 负类样本集$D_{-1}=\{x_3=(-1,-3)^T, x_4=(-3,-2)^T\}$.

对数据实例$x=(0,0)^T$来说,

- 如果我们选择曼哈顿距离$dist_1$, 则

$$dist_1(x_1,x)=4,\ dist_1(x_2,x)=4,$$

$$dist_1(x_3,x)=4,\ dist_1(x_4,x)=5.$$

于是与x最邻近的3个点构成的x的邻域

$$N_3^{dist_1}(x)=\{x_1,x_2,x_3\},$$

则我们将x的类别预测为$y=+1$.

距离度量对预测结果的影响

对数据实例$x = (0, 0)^T$来说，

- 如果我们选择欧氏距离$dist_2$，则

$$dist_2(x_1, x) = 2\sqrt{2},\ dist_2(x_2, x) = 4,$$

$$dist_1(x_3, x) = \sqrt{10},\ dist_1(x_4, x) = \sqrt{13}.$$

于是与x最邻近的3个点构成的x的邻域

$$N_3^{dist_2}(x) = \{x_1, x_3, x_4\},$$

则我们将x的类别预测为$y = -1$.

距离度量对预测结果的影响

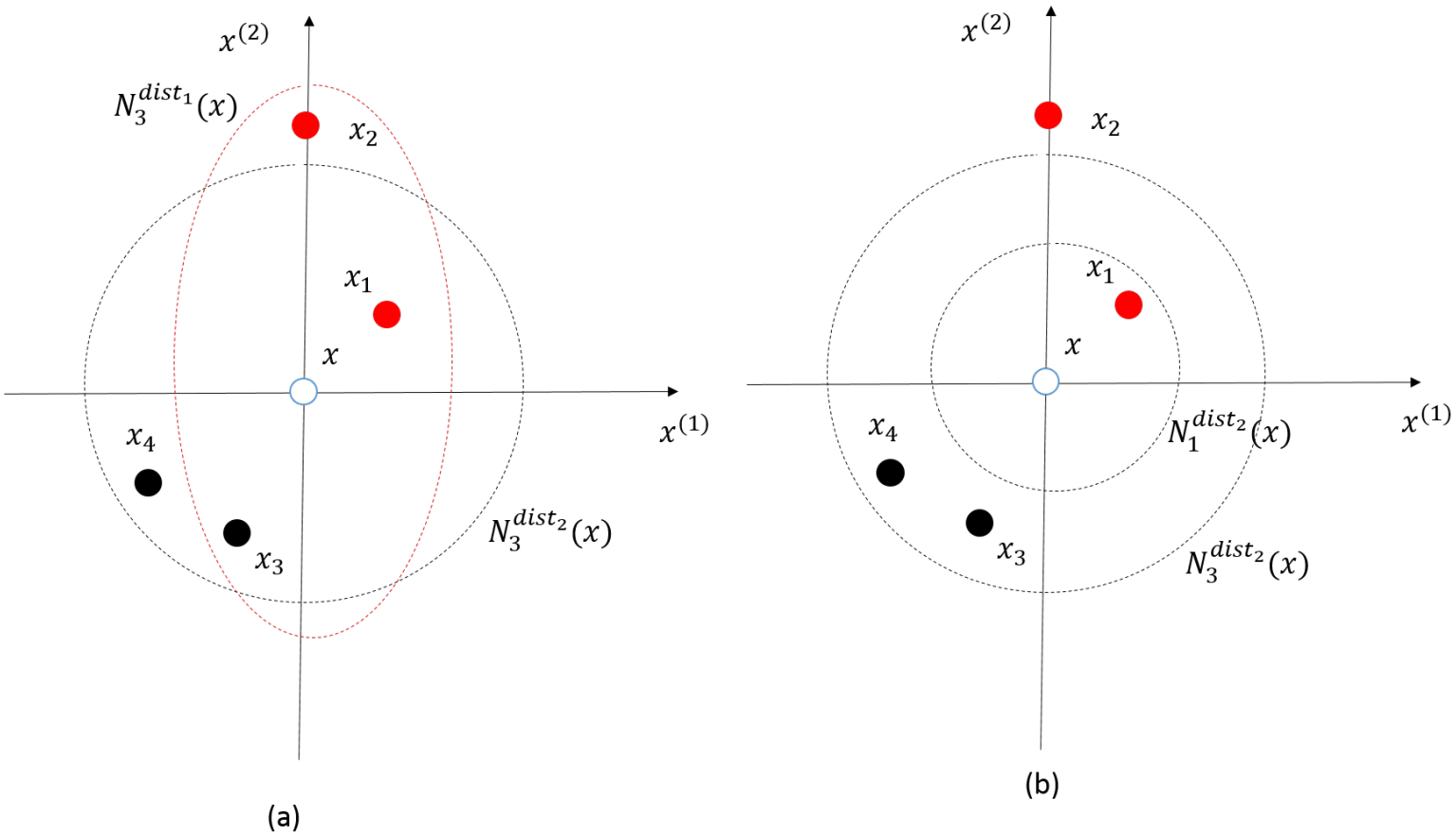

Figure: 距离选择与k值选择对预测结果的影响

k值选择对预测结果的影响

对数据实例$x = (0, 0)^T$来说，如果我们选择欧氏距离$dist_2$，且

- 令$k = 3$，则与x最邻近的3个点构成的x的邻域

$$N_3^{dist_2}(x) = \{x_1, x_3, x_4\},$$

则我们将x的类别预测为$y = -1$;

- 令$k = 1$，则与x最邻近的样本点为x_1，即

$$N_1^{dist_2}(x) = \{x_1\},$$

此时x的类别预测为$y = +1$.

k值选择对预测结果的影响

如何选择合适的 k值呢？

- 如果选择比较大的k值，与输入实例不太相似的训练样本也可能会对输入实例的类标签预测起作用，导致预测错误；
- 但如果k值较小，则预测结果依赖于个别和输入实例相对很相似的训练样本，容易导致预测对噪声扰动的容忍程度不高.
- 一般先从比较小的k 值开始，依次增加，选择在验证集上分类错误率最小的k值.

概要

最近邻法

- 对k-近邻法来说，如果$k = 1$，则称为最近邻法.
- 最近邻法对输入实例点x的类别预测完全由与x最近的训练样本的类别标记确定，即

$$y = \underset{y_i}{\operatorname{argmin}}\, dist(x_i, x).$$

- 这使得最近邻法的预测偏差小而方差大.

- 如果对任意输入实例x附近任意小的距离δ范围内都总能找到一个训练样本(这当然要求样本量足够大)，则我们可以认为x的最近邻点z和x的特征基本一样，进而认为

$$P(y_z = c_k|z) = P(y = c_k|x),$$

其中y_z为z 的类标记。

- 最近邻法的泛化错误率ϵ为x和其最近邻点属于不同类别的概率，则在上述条件下ϵ可以表示成

$$\epsilon = \sum_{m=1}^{M} P(y = c_m|x)(1 - P(y = c_m|x)).$$

- 令$c_{m^*} = \underset{c_m}{\operatorname{argmax}} P(y = c_m|x)$, 则贝叶斯错误率$\epsilon_B$为

$$\epsilon_B = 1 - P(y = c_{m^*}|x).$$

- Cover和Hart证明了

$$\epsilon_B \leq \epsilon \leq 2\epsilon_B - \frac{M}{M-1}\epsilon_B^2,$$

即 最近邻法的泛化错误率ϵ 不超过贝叶斯错误率ϵ_B的两倍.

- 在高维特征空间中，要求任意输入实例附近任意小的距离范围内都总能找到一个训练样本比较困难.
- 最近邻法通常适合低维特征空间的分类任务.

特征空间划分

- 最近邻法其实将特征空间进行了一个划分$\mathcal{X} = \bigcup_{i=1}^{N} R_i$.
- 对每个划分单元R_i来说，该单元的数据点到其他训练样本的距离都不会小于到样本x_i的距离，即

$$R_i = \{x \in \mathcal{X} | dist(x, x_i) = \min_{1 \le j \le N} dist(x, x_j)\}.$$

- 落入每个划分单元R_i的测试样本的类别预测和该单元相对应的训练样本x_i的类别标记保持一致，即对任一$x \in R_i$，都有$y = y_i$.
- 从这个角度来说，每个样本点对相应的划分单元提供了一个表示.

特征空间划分

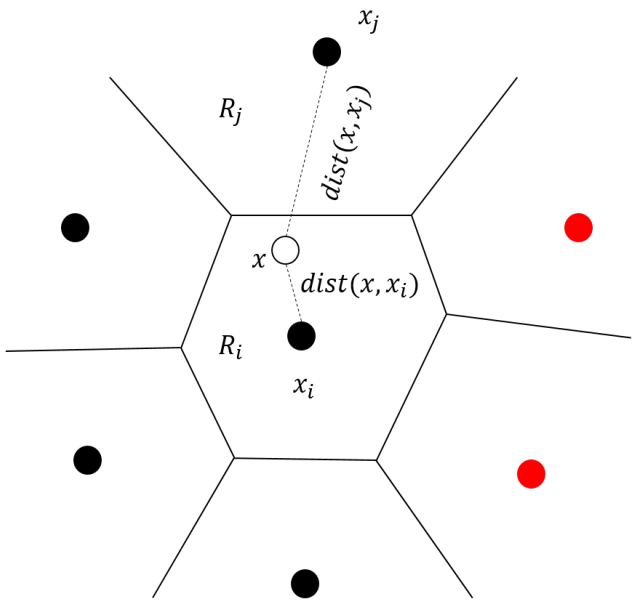

Figure: 最近邻法对空间的划分

概要

- 如果训练样本集中存在一组同类别的样本点非常相似，直观上这些非常相似的样本点所代表的划分单元中的数据点应该也比较相似.
- 能否将这组相似的样本点对应的划分单元合在一起看作一个比较大的数据区域?
- 如何处理这一组非常相似的代表点?
- 能否通过一个特定的数据点(比如这组样本点的中心等)来替代这组训练样本点作为这个数据区域的代表，从而得到一个规模比较小的新“训练数据集”?
- 如果能有这样的替换，对新数据实例类别的预测任务就可以考虑在这样一个较小规模的替代训练数据集上应用最近邻法.

要实现这样的想法，需要解决两个问题：

- 首先需要确定训练样本集的划分机制:能够将训练样本集D划分成若干互斥的子集使得
  - 每个子集中的训练样本点尽可能相互比较相似，
  - 不同子集的样本尽可能不同.
- 其次需要设定用以确定每个子集的代表点的机制.

基于K-means的分类方法

给定训练样本集$D = \{(x_i, y_i)\}_{i=1}^{N}$,以$D_i$表示属于类$c_i$的训练样本集，$i = 1, 2, \cdots, M$.

- 采用K-means方法将每个D_i划分为K个单元.
- 以属于每个单元的训练样本的特征向量的均值作为该单元的代表.

那么K-means方法如何将每个D_i划分为K个单元呢？我们不妨设

$$(D_{i1}, D_{i2}, \cdots, D_{iK})$$

为D_i的一个划分，则每个D_{ij}中样本特征向量的均值为

$$c_{ij} = \frac{1}{|D_{ij}|} \sum_{(x_t, y_t) \in D_{ij}} x_t.$$

K-means方法的目标是寻找一个划分$(D_{i1}^*, D_{i2}^*, \cdots, D_{iK}^*)$使得数据分布的方差最小，即

$$(D_{i1}^*, D_{i2}^*, \cdots, D_{iK}^*) = \underset{D_{i1}, D_{i2}, \cdots, D_{iK}}{\operatorname{argmin}} \sum_{j=1}^{K} \sum_{(x_t, y_t) \in D_{ij}} \| x_t - c_{ij} \|_2^2$$

K-means方法

方差最小的划分通常在实际应用中很难达到，一般采用如下的近似划分方法：

(1) 选择K个点$c_{i1}, c_{i2}, \cdots, c_{iK}$作为初始点；

(2) 按如下方法构造划分$D_{i1}, D_{i2}, \cdots, D_{iK}$：对每个$(x_t, y_t) \in D_i$(这里$y_t = c_i$)，令

$$l_{x_t} = \underset{j}{\operatorname{argmin}} \parallel x_t - c_{ij} \parallel_2,$$

则

$$D_{ij} = \{(x_t, y_t) \in D_i \wedge l_{x_t} = j\}$$

即将x_t划分到和x_t 距离最近的均值点对应的单元中；

K-means方法

(3) 对每个D_{ij}，更新其均值为

$$c_{ij} = \frac{1}{|D_{ij}|} \sum_{(x_t, y_t) \in D_{ij}} x_t.$$

(4) 重复(2)和(3)，直到收敛;

(5) 返回划分$D_{i1}, D_{i2}, \cdots, D_{iK}$及其相应的均值向量$c_{i1}, c_{i2}, \cdots, c_{iK}$.

基于K-means的分类方法

- 我们以$c_{i1}, c_{i2}, \cdots, c_{iK}$作为$D_i$的划分$D_{i1}, D_{i2}, \cdots, D_{iK}$的代表点.
- 并且以c_i作为每个c_{ij}的类标记.
- 进一步， 我们可以将$M \times K$个代表点构成的点集

$$D' = \{(c_{i1}, c_i), (c_{i2}, c_i), \cdots, (c_{iK}, c_i)\}_{i=1}^{M}$$

作为“训练样本集”.

- 对新输入数据点x采用最近邻法进行分类，将与x最近的代表点的类标记作为x的类标记， 即

$$y = \underset{c_i}{\operatorname{argmin}} \min_{1 \le k \le K} dist(x, c_{ik}).$$

基于K-means的分类方法

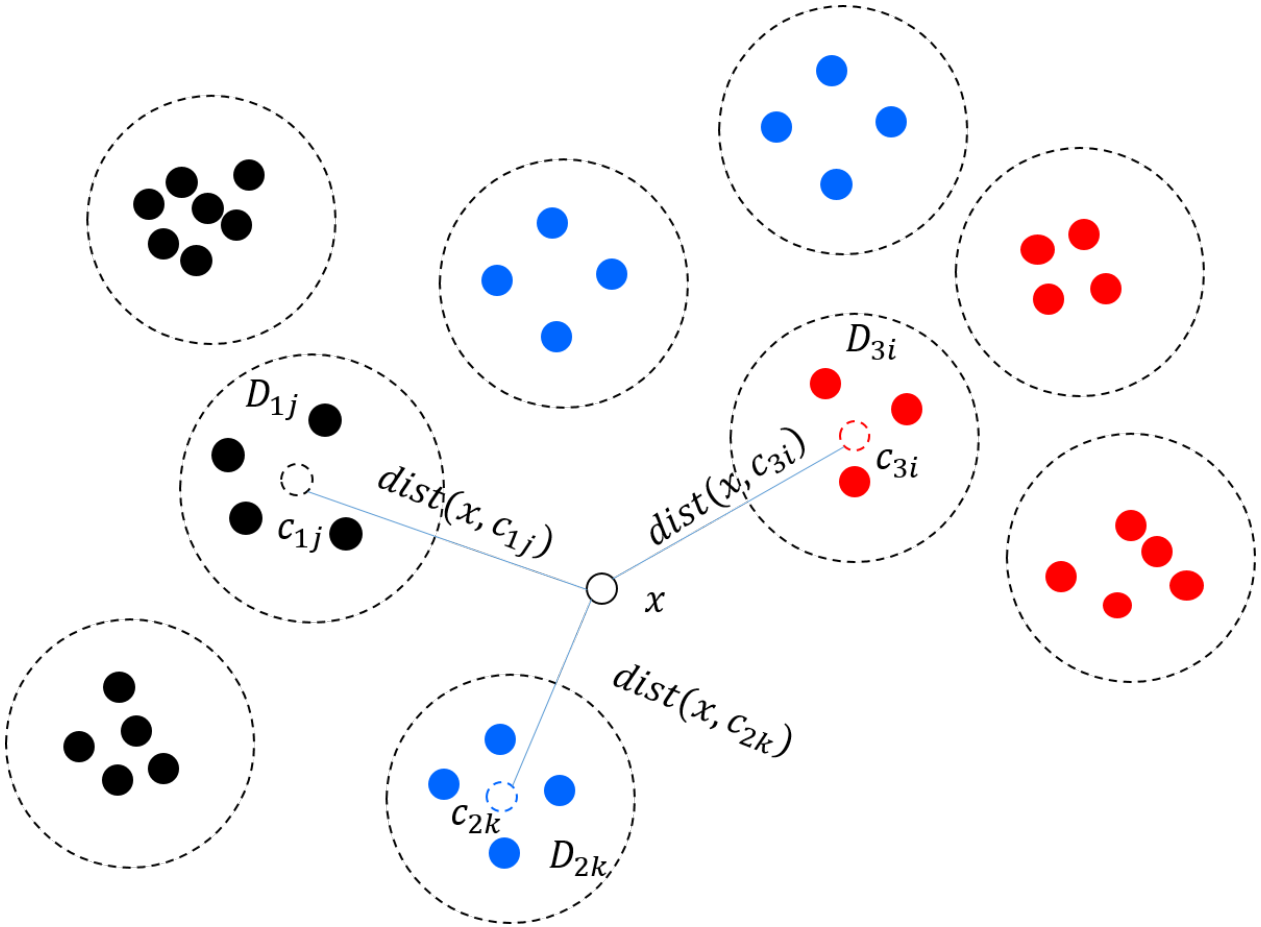

Figure: $M = K = 3$时基于K-means的最近邻法

- 基于K-means的分类方法只是基于同类样本来独立确定区域代表点.
- 可能会使得某些代表点离分类边界比较近而导致使用最近邻原则分类时出现误分类的情况.

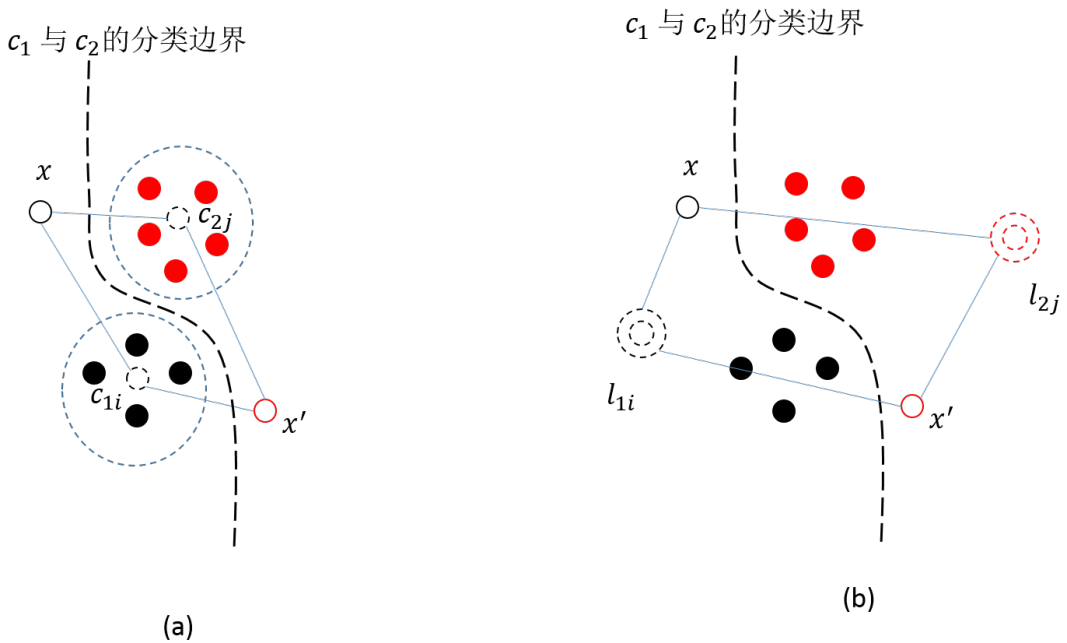

Figure: K-means与学习向量量化确定的代表点之比较

学习向量量化方法

- 学习向量量化方法对每个类c_m构建K个代表点.
- 遵循“代表点靠近同类训练样本而背离异类样本”的原则来调整代表点，使得同类和异类样本都在代表点的构建过程中起作用.

学习向量量化算法框架：

(1) 对每个类c_m随机选择K个点

$$l_{m1}, l_{m2}, \cdots, l_{mK}$$

作为其代表点向量的初始值，并以c_m作为这些代表点的类标记；

(2) 随机选择一个训练样本$(x_i, y_i) \in D$，找出与x_i最近的代表点$l_{m^*k^*}$，即

$$l_{m^*k^*} = \underset{l_{mk}}{\operatorname{argmin}} \parallel x_i - l_{mk} \parallel_2$$

学习向量量化算法框架

(3) 如果$y_i = m^*$，则对$l_{m^*k^*}$进行如下更新：

$$l_{m^*k^*} \leftarrow l_{m^*k^*} + \eta(x_i - l_{m^*k^*})$$

否则对$l_{m^*k^*}$进行如下更新：

$$l_{m^*k^*} \leftarrow l_{m^*k^*} - \eta(x_i - l_{m^*k^*})$$

(4) 重复(2)(3)，直到满足停至条件。

这里$\eta \in (0, 1)$是学习率。

- 和基于K-means的分类方法一样，可以将代表点的集合视为新的“训练样本集”，而对新输入的测试数据应用最近邻法来预测其类标记。

在代表点更新中，如果x_i和是$l_{m^*l^*}$是同类别的，则

$$\begin{aligned} & \| x_i - [l_{m^*k^*} + \eta(x_i - l_{m^*k^*})] \|_2 \\ = \quad & (1-\eta) \| x_i - l_{m^*k^*} \|_2 \\ \leq \quad & \| x_i - l_{m^*k^*} \|_2; \end{aligned}$$

反之，则

$$\begin{aligned} & \| x_i - [l_{m^*k^*} - \eta(x_i - l_{m^*k^*})] \|_2 \\ = \quad & (1+\eta) \| x_i - l_{m^*k^*} \|_2 \\ \geq \quad & \| x_i - l_{m^*k^*} \|_2 . \end{aligned}$$

小结

- k-近邻法
- 最近邻法
- 最近邻的扩展
  - 基于K-means的分类方法
  - 学习向量量化方法
- 最近邻法及其扩展方法中的代表点也可以被称为相应数据区域或者单元的原型（Prototype），因此这类方法通常被称作原型方法或者免模型(model-free)方法.