Machine Learning

Yuh-Jye Lee

Data Science and Machine Intelligence Lab
National Chiao Tung University

March 7, 2017

Instance-based Learning

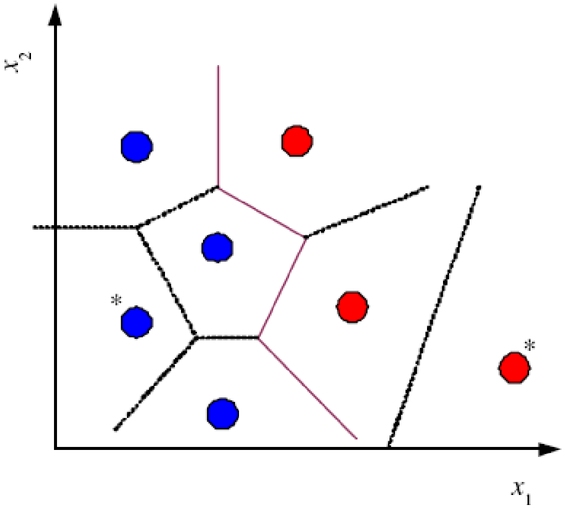
- Fundamental philosophy: Two instances that are *close to each other* or *similar to each other* they should share with the same label
- Also known as *memory-based learning* since what they do is store the training instances in a lookup table and *interpolate* from these.
- It requires memory of $\mathcal{O}(N)$
- Given an input similar ones should be found and finding them requires computation of $\mathcal{O}(N)$
- Such methods are also called *lazy learning* algorithms. Because they do NOT compute a model when they are given a training set but postpone the computation of the model until they are given a new test instance (query point)

k-Nearest Neighbors Classifier

- Given a query point x^o, we find the k training points $x^{(i)},\ i = 1, 2, \dots, k$ *closest* in *distance* to x^o
- Then classify using *majority vote* among these k neighbors.
- Choose k as an odd number to avoid the *tie*. Ties are broken at random
- If all attributes (features) are real-valued, we can use Euclidean distance. That is $d(x, x^o) = \|x - x^o\|_2$
- If the attribute values are *discrete*, we can use *Hamming distance*, which counts the number of *nonmatching* attributes

$$d(x, x^o) = \sum_{j=1}^{n} \mathbf{1}(x_j \neq x_j^o) \tag{1}$$

1-Nearest Neighbor Decision Boundary



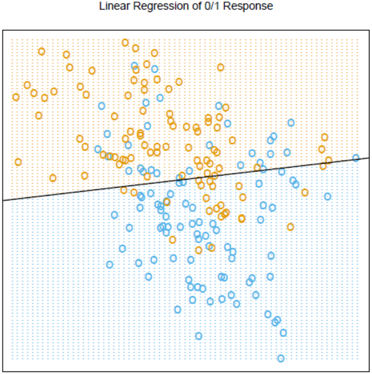


Figure: A classification example in two dimensions. The classes are coded as a binary variable (BLUE=0, ORANGE=1), and then fit by linear regression. The line is the decision boundary defined by $\mathbf{x}^T\hat{\beta} = 0.5$. The orange shaded region denotes that part of input space classified as ORANGE, while the blue region is classified as Blue. From "The Elements of Statistical Learning, Hastie et al."

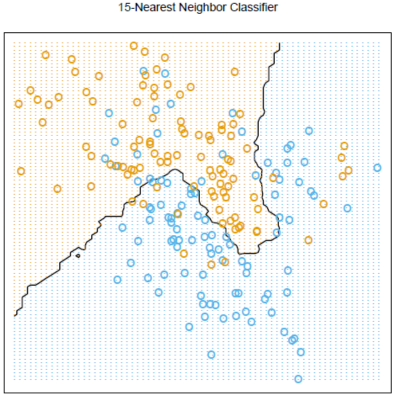


Figure: The same classification example in two dimensions as in the previous figure . The classes are coded as a binary variable (BLUE=0, ORANGE=1) and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors. From "The Elements of Statistical Learning, Hastie et al."

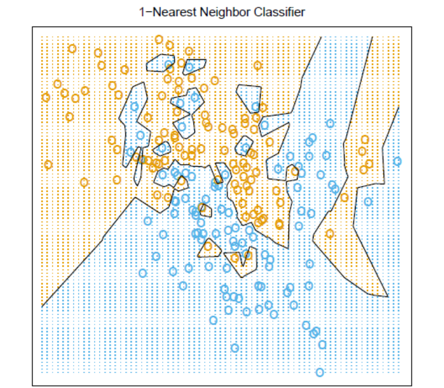
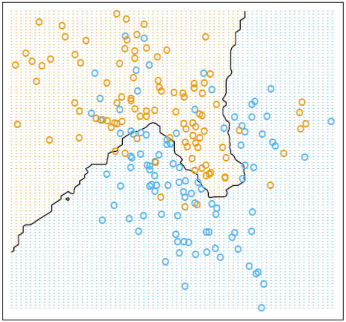
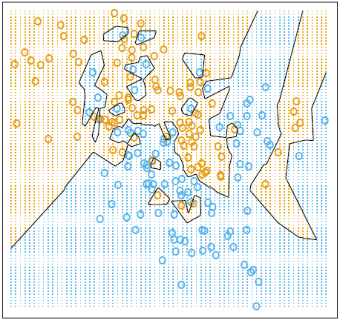


Figure: The same classification example in two dimensions as in the previous figure . The classes are coded as a binary variable (BLUE=0, ORANGE=1), and then predicted by 1-nearest-neighbor classification. From "The Elements of Statistical Learning, Hastie et al."





(a) 15-Nearest Neighbor Classifier (b) 1-Nearest Neighbor Classifier

Figure: The Comparison between 15-Nearest Neighbor Classifier and 1-Nearest Neighbor Classifier. From "The Elements of Statistical Learning, Hastie et al."

Condensed Nearest Neighbor

$\mathcal{Z} \leftarrow \emptyset$

Repeat

  For all $\boldsymbol{x} \in \mathcal{X}$ (in random order)

    Find $\boldsymbol{x}' \in \mathcal{Z}$ s.t. $\|\boldsymbol{x} - \boldsymbol{x}'\| = \min_{\boldsymbol{x}^j \in \mathcal{Z}} \|\boldsymbol{x} - \boldsymbol{x}^j\|$

    If class($\boldsymbol{x}$)$\neq$class($\boldsymbol{x}'$) add $\boldsymbol{x}$ to $\mathcal{Z}$

Until $\mathcal{Z}$ does not change

Distance Measure

- Using different distance measurements will give very different results in k-NN algorithm.
- Be careful when you compute the distance
- We might need to *normalize* the scale between different attributes. For example, yearly income vs. daily spend
- Typically we first standardize each of the attributes to have mean zero and variance 1

$$\hat{x}_j = \frac{x_j - \mu_j}{\sigma_j} \tag{2}$$

Learning Distance Measure

- Finding a distance function $d(x^i, x^j)$ such that if x^i and x^j are belong to the *class* the distance is *small* and if they are belong to the *different classes* the distance is large.
- Euclidean distance: $\|x^i - x^j\|_2^2 = (x^i - x^j)^\top (x^i - x^j)$
- Mahalanobis distance: $d(x^i, x^j) = (x^i - x^j)^\top M (x^i - x^j)$ where M is a positive semi-definited matrix.

$$
\begin{aligned}
&(x^i - x^j)^\top M (x^i - x^j) \\
&= (x^i - x^j)^\top L^\top L (x^i - x^j) \\
&= (Lx^i - Lx^j)^\top (Lx^i - Lx^j)
\end{aligned}
$$

- The matrix L can be with the size $k \times n$ and $k << n$