Machine Learning (CS 5350/CS 6350) *22 Jan 2008*

K-nearest-neighbors

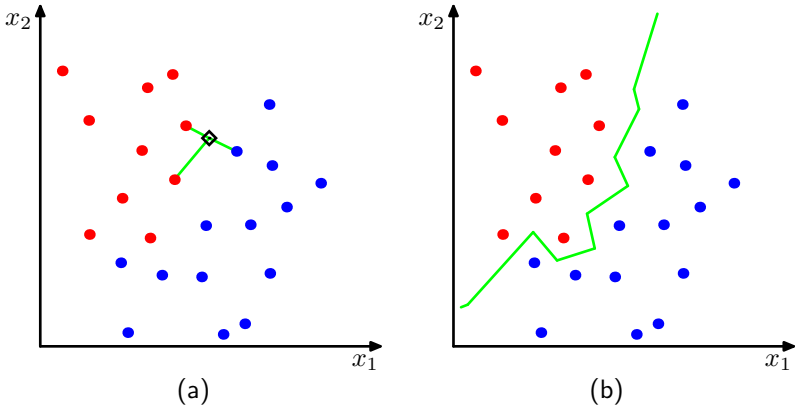
1NN—simple intuition: to classify a new data point, just return the class of the closest training point.

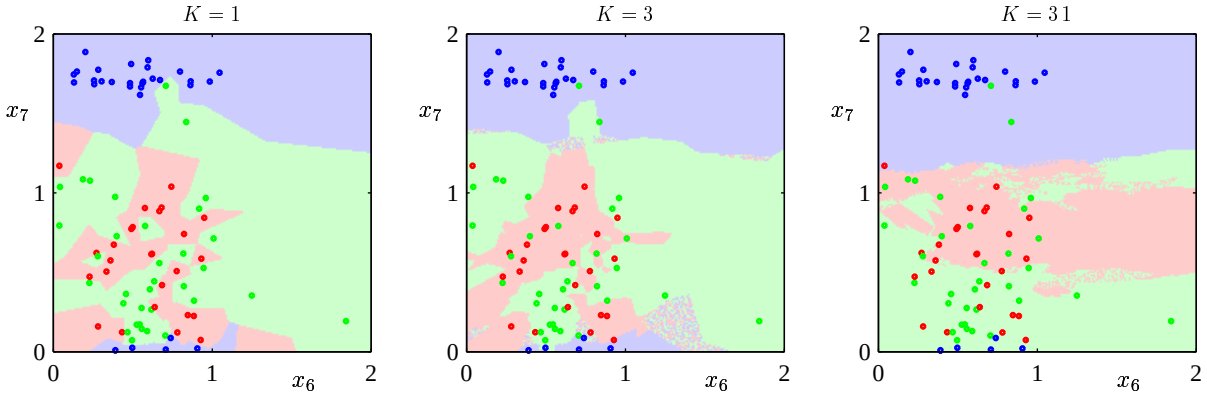
kNN—instead of single closest point, average over the k nearest.

ϵNN—instead of the k nearest, use as many as fit in a ball of radius ϵ.

How does one train such an algorithm?

Despite its simplicity, kNN is a really really good classifier. (But very sensitive to irrelevant features.)





In general, as k (or ϵ) gets bigger, you are risking more underfitting. As k (or ϵ) gets smaller, you are risking more overfitting.

One can also give each of the k nearest neighbors a *vote* based on how far away they are. This vote is usually of the form $1/d$ or e^{-d}, where d is the distance between the training point and the test point.

(Can be used also for regression, by averaging responses.)