Classification: k-Nearest Neighbor & Instance-based Learning

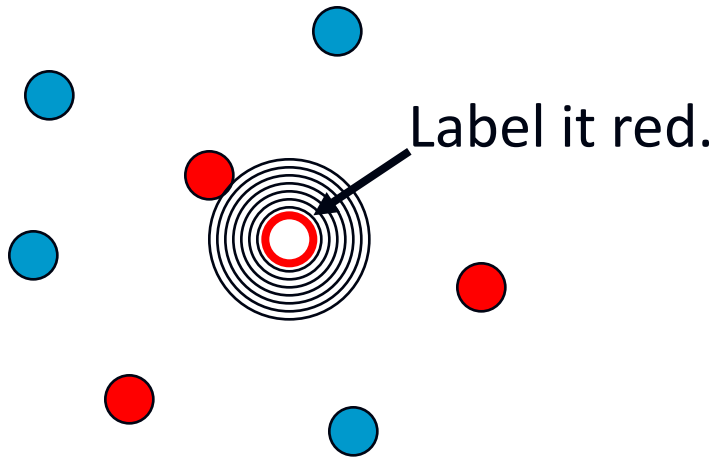
Some material adapted from slides by Andrew Moore, CMU.

Visit http://www.autonlab.org/tutorials/ for
Andrew's repository of Data Mining tutorials.

These slides were assembled by Byron boots based on the slides assembled by Eric Eaton, with grateful acknowledgement of the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution.

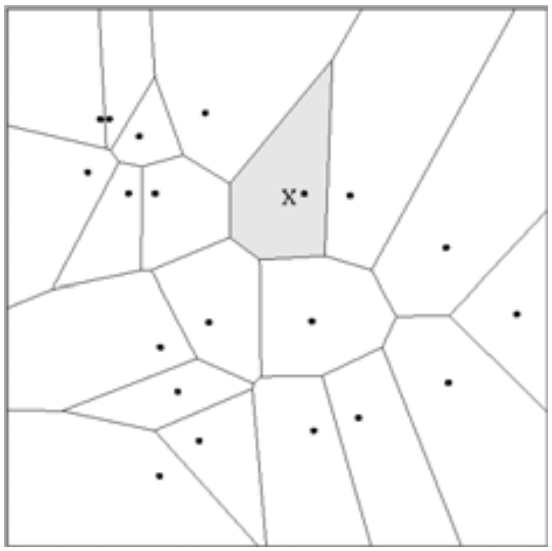
1-Nearest Neighbor

- One of the simplest of all machine learning classifiers
- Simple idea: label a new point the same as the closest known point



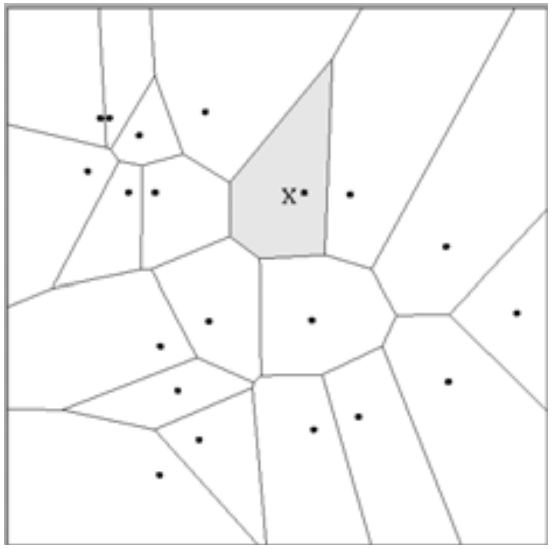
1-Nearest Neighbor

- A type of instance-based learning
  - Also known as "memory-based" learning
- Forms a Voronoi tessellation of the instance space

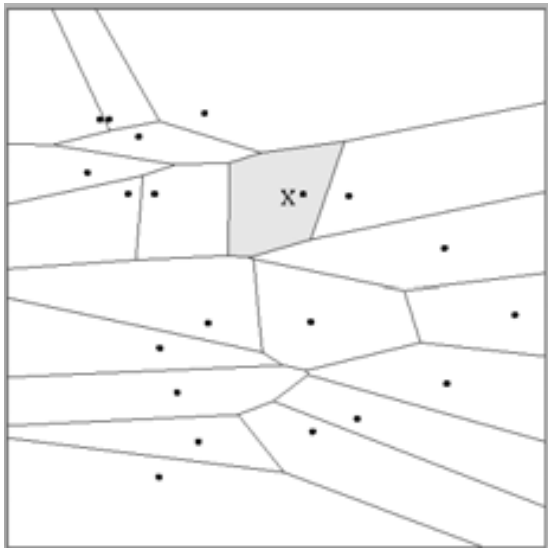


Distance Metrics

- Different metrics can change the decision surface



Dist(**a**,**b**) =$(a_1 - b_1)^2 + (a_2 - b_2)^2$



Dist(**a**,**b**) =$(a_1 - b_1)^2 + (3a_2 - 3b_2)^2$

- Standard Euclidean distance metric:
  - Two-dimensional: Dist(a,b) = sqrt($(a_1 - b_1)^2 + (a_2 - b_2)^2$)
  - Multivariate: Dist(a,b) = sqrt($\sum (a_i - b_i)^2$)

Adapted from "Instance-Based Learning" lecture slides by Andrew Moore, CMU.

Four Aspects of an Instance-Based Learner:

1. A distance metric
2. How many nearby neighbors to look at?
3. A weighting function (optional)
4. How to fit with the local points?

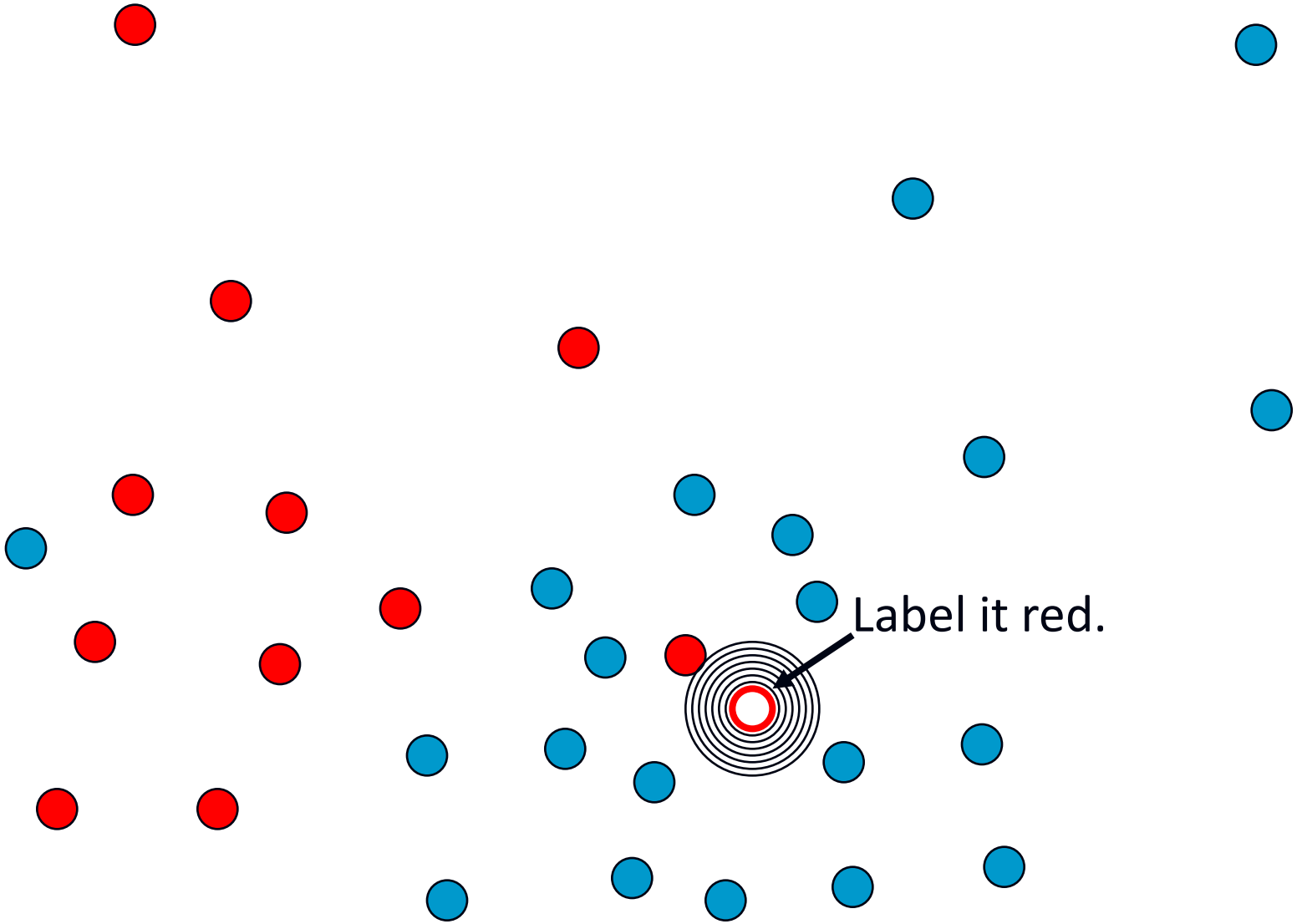
Adapted from "Instance-Based Learning" lecture slides by Andrew Moore, CMU.

1-NN's Four Aspects as an Instance-Based Learner:

1. A distance metric
   - *Euclidian*
2. How many nearby neighbors to look at?
   - *One*
3. A weighting function (optional)
   - *Unused*
4. How to fit with the local points?
   - *Just predict the same output as the nearest neighbor.*

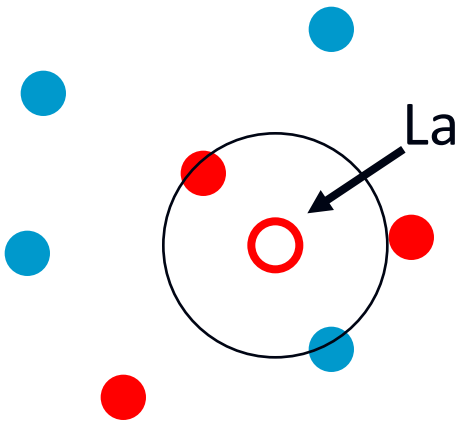
Adapted from "Instance-Based Learning" lecture slides by Andrew Moore, CMU.

1-Nearest Neighbor



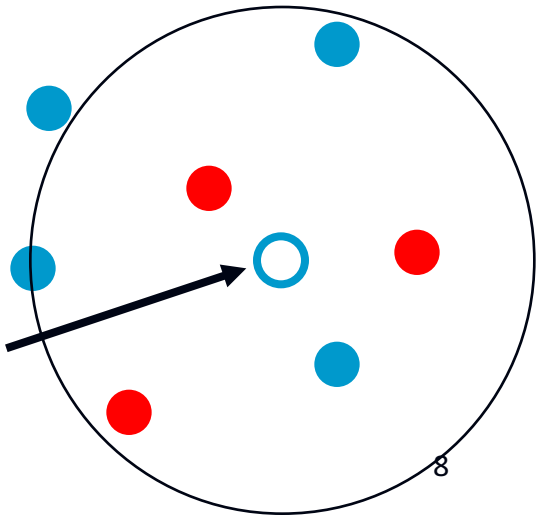
k – Nearest Neighbor

- Generalizes 1-NN to smooth away noise in the labels
- A new point is now assigned the most frequent label of its *k* nearest neighbors

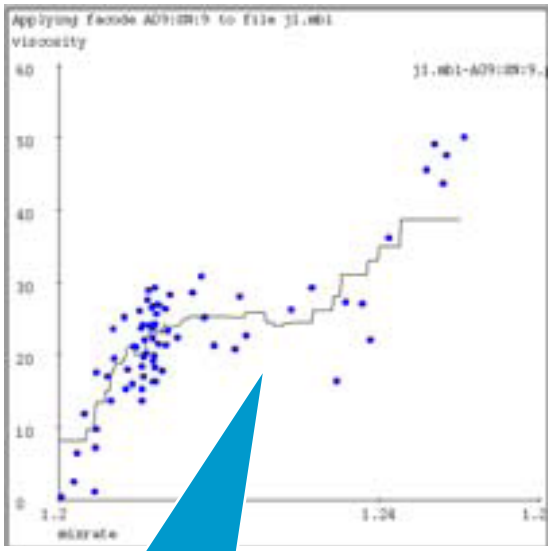


Label it red, when k = 3

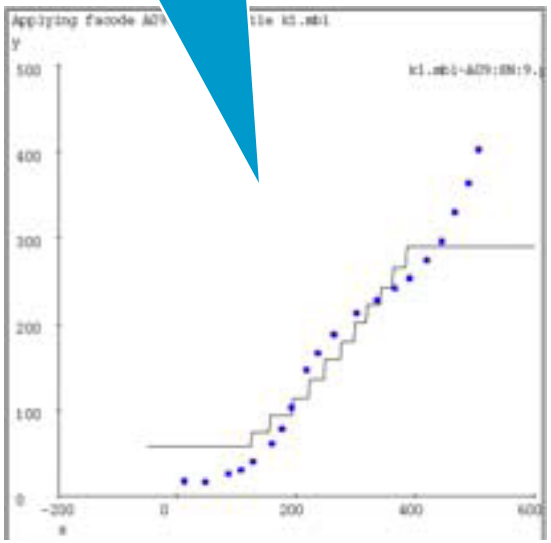
Label it blue, when k = 7



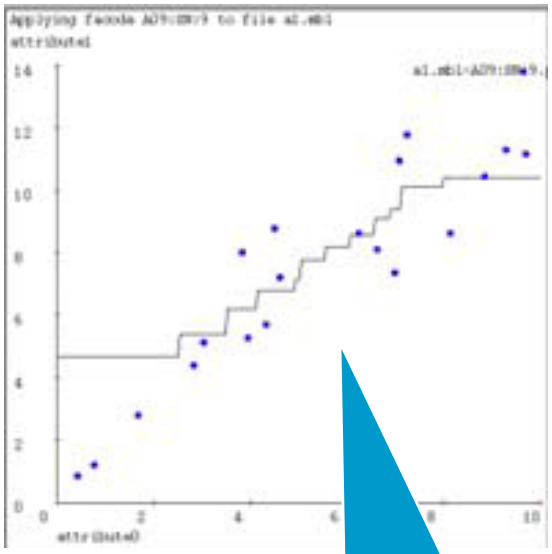
k-Nearest Neighbor (k = 9)



A magnificent job of noise smoothing. Three cheers for 9-nearest-neighbor.
...But the jerkiness isn't good.



Appalling behavior! Loses all the detail that 1-nearest neighbor would give. The tails are horrible!



Fits much less of the noise, captures trends. But still, frankly, pathetic compared with linear regression.

Adapted from "Instance-Based Learning" lecture slides by Andrew Moore, CMU.