Introduction to Machine Learning

Geometry and Nearest Neighbors

Geometry and dimensionality

Engineering feature vectors

From: Maria Dieudonne (mairamdidonne@gmail.com)

Subject: HELLO

Dearest, I hope you received my message last week about the estate ? Please I am still waiting for your reply to proceed with you,am sick now but will introduce you to my son to proceed with you on the project.Send me your direct number to call you for oral discussion. Remain blessed Mrs.Tereza Dijiriga

From: Voiland Communications (VCEA.News@wsu.edu)

Subject: VCEA Town Hall w/ WSU's President Kirk Schulz and Provost Dan Bernardo

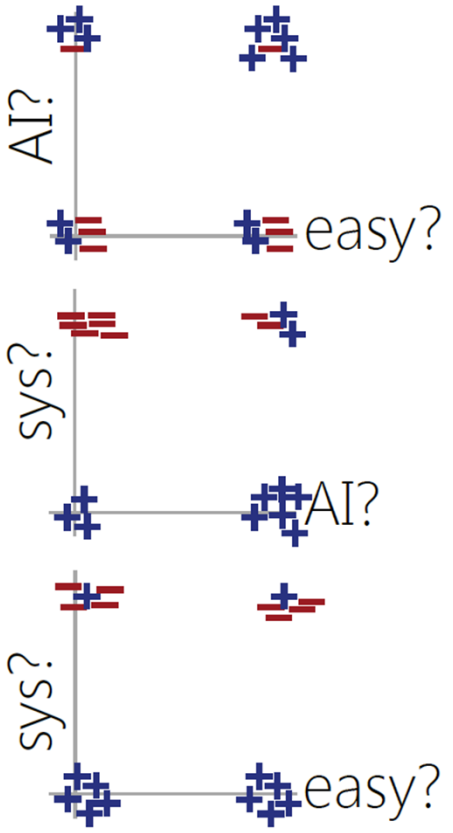
Faculty, staff, and students from throughout the WSU system are invited to join President Kirk Schulz and Provost Dan Bernardo for Town Halls 2018, a series of informal meetings scheduled this fall for each campus, college, and major administrative unit.

The get-togethers will include a brief presentation and a Q&A session.

Topics
- Drive to 25: Action plans and progress
- Other top priorities
- Recent notable achievements
- Your questions

Features as dimensions

Rating	Easy?	AI?	Sys?	Thy?	Morning?
+2	y	y	n	y	n
+2	y	y	n	y	n
+2	n	y	n	n	n
+2	n	n	n	y	n
+2	n	y	y	n	y
+1	y	y	n	n	n
+1	y	y	n	y	n
+1	n	y	n	y	n
0	n	n	n	n	y
0	y	n	n	y	y
0	n	y	n	y	n
0	y	y	y	y	y
-1	y	y	y	n	y
-1	n	n	y	y	n
-1	n	n	y	n	y
-1	y	n	y	n	y
-2	n	n	y	y	n
-2	n	y	y	n	y
-2	y	n	y	n	n
-2	y	n	y	n	y



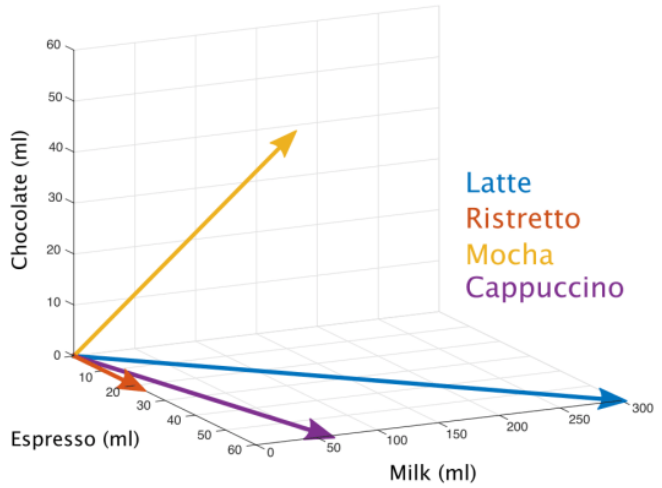
Mapping feature values to vectors

- Systems
  - Yes → 1
  - No → 0

- Color
  - Red →
  - Green →
  - Blue →
  - Yellow →
  - Black →
  - White →

Instance as vector in high-dimensional feature space

- $x = \langle x_1, x_2, .., x_D \rangle$
- D-dimensional space

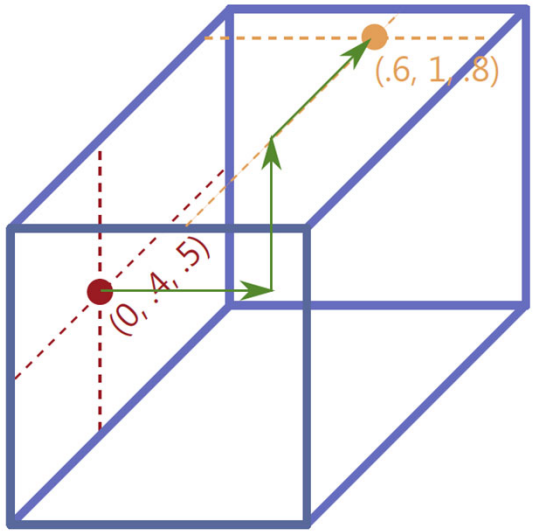


K-Nearest neighbors (knn) classifier

- Inductive bias: Label for an instance should be similar to the label of nearby points
- Label for test example $\hat{x}$ predicted to be the same as its closest neighbors

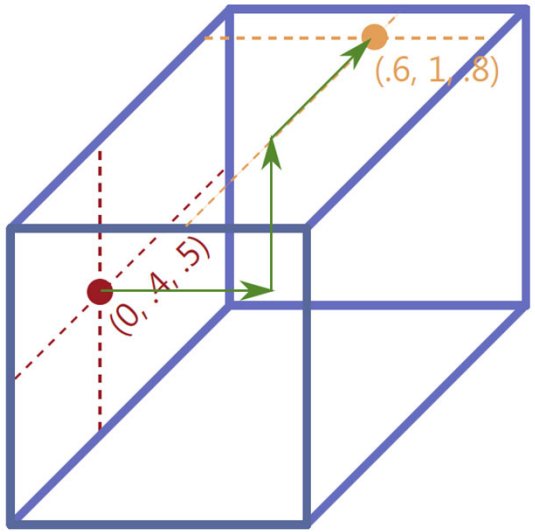
K-Nearest Neighbors classifier

- Computes distance between instances in vector space



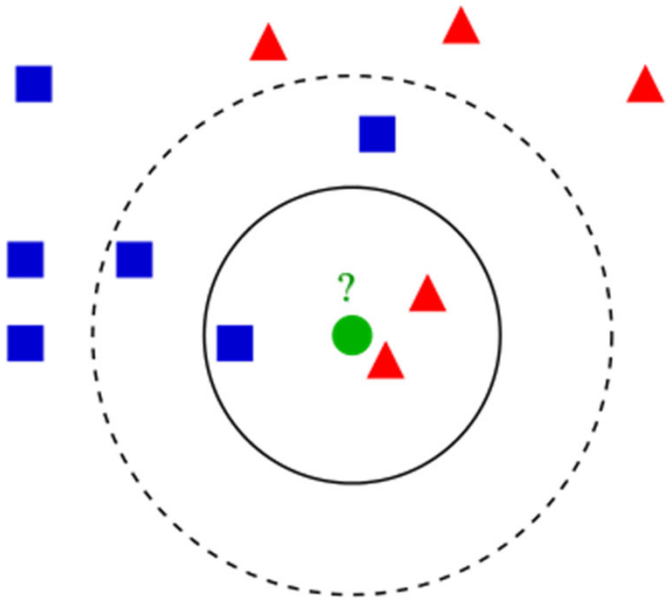
Euclidean distance

$$d(a,b) = \left[\sum_{d=1}^{D}(a_d - b_d)^2\right]^{\frac{1}{2}}$$

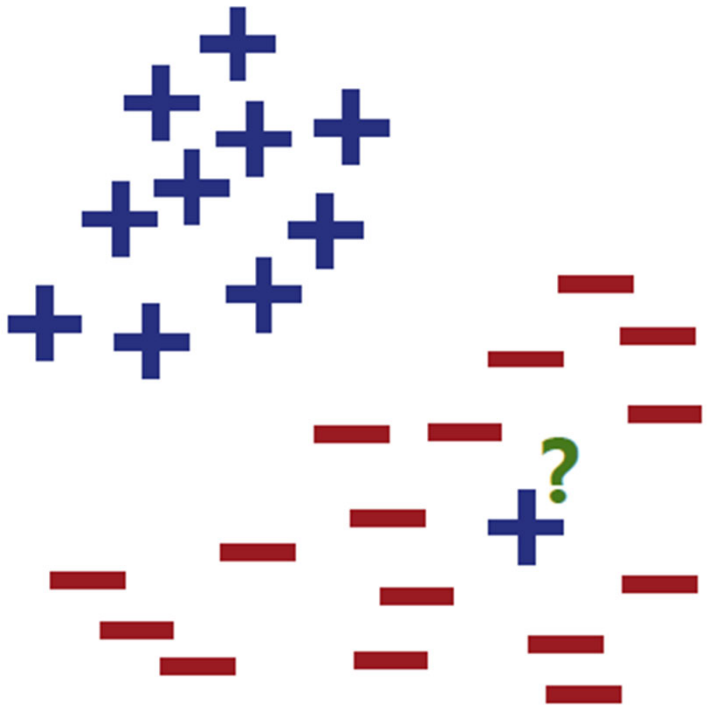


Classification

- Training data ***D***
- Training examples $(x_1,y_1),..,(x_N,y_N)$



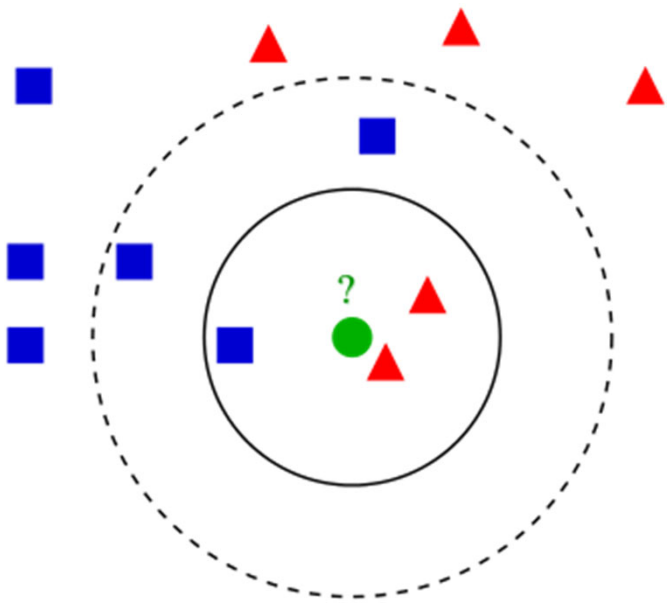
Nearest neighbor



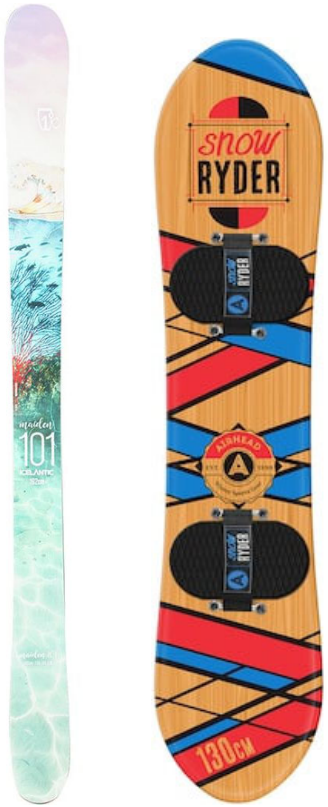
Algorithm 3 KNN-PREDICT($\mathbf{D}$, K, $\hat{x}$)

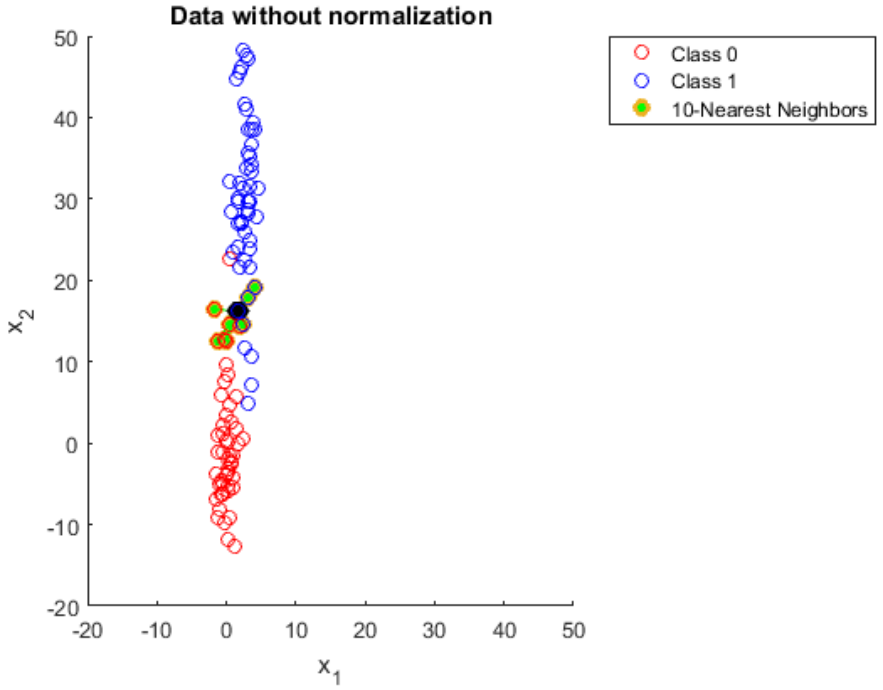
$S \leftarrow [\,]$
for $n = 1$ **to** N **do**
  $S \leftarrow S \oplus \langle \mathrm{d}(x_n, \hat{x}), n \rangle$ // store distance to training example n
end for
$S \leftarrow \textsc{sort}(S)$ // put lowest-distance objects first
$\hat{y} \leftarrow 0$
for $k = 1$ **to** K **do**
  $\langle dist, n \rangle \leftarrow S_k$ // n this is the kth closest data point
  $\hat{y} \leftarrow \hat{y} + y_n$ // vote according to the label for the nth training point
end for
return $\textsc{sign}(\hat{y})$ // return $+1$ if $\hat{y} > 0$ and -1 if $\hat{y} < 0$

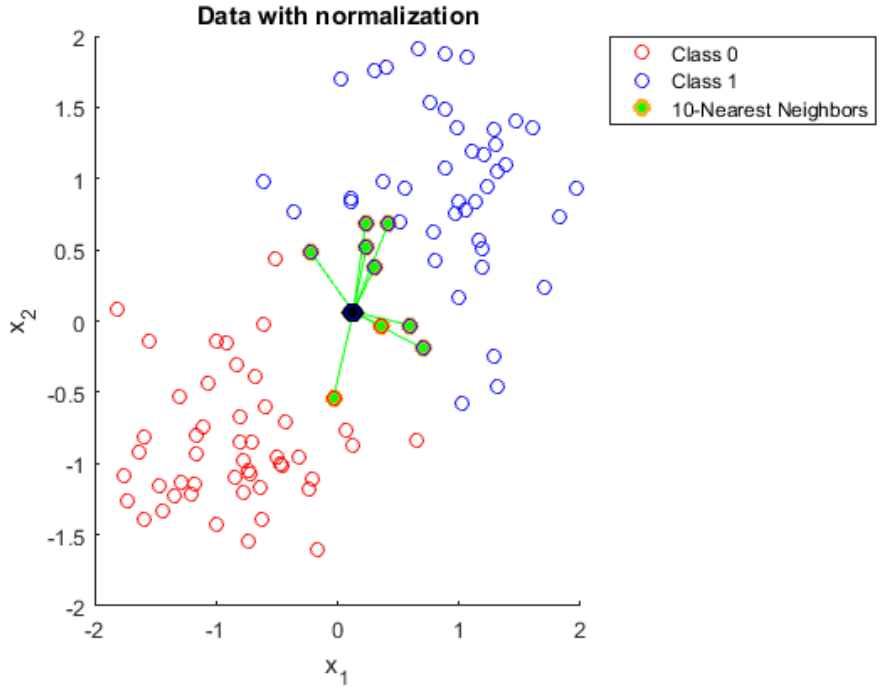
How do we choose k?



When can a distance function cause the classifier to fail?

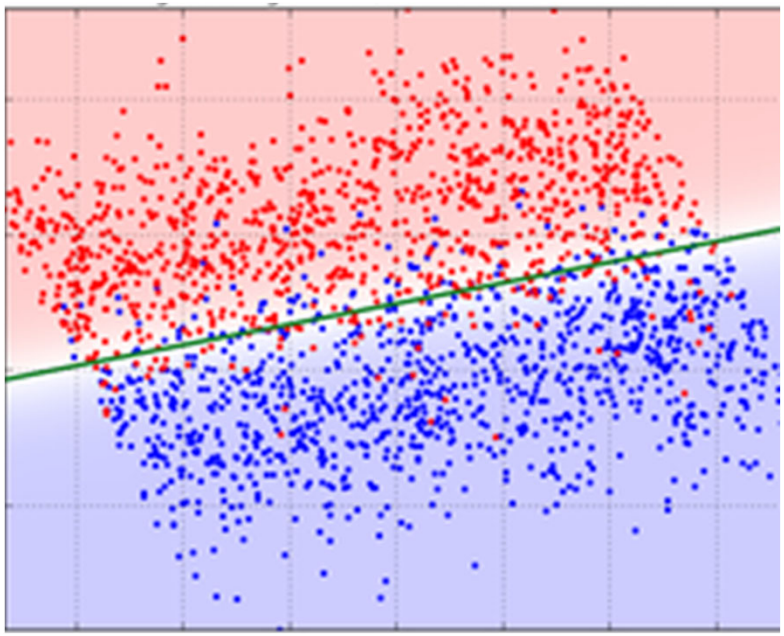



Data without normalization
Class 0
Class 1
10-Nearest Neighbors
x_1
x_2
50
40
30
20
10
0
-10
-20
-20
-10
0
10
20
30
40
50


Data with normalization
Class 0
Class 1
10-Nearest Neighbors
x_1
x_2
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1
0
1
2

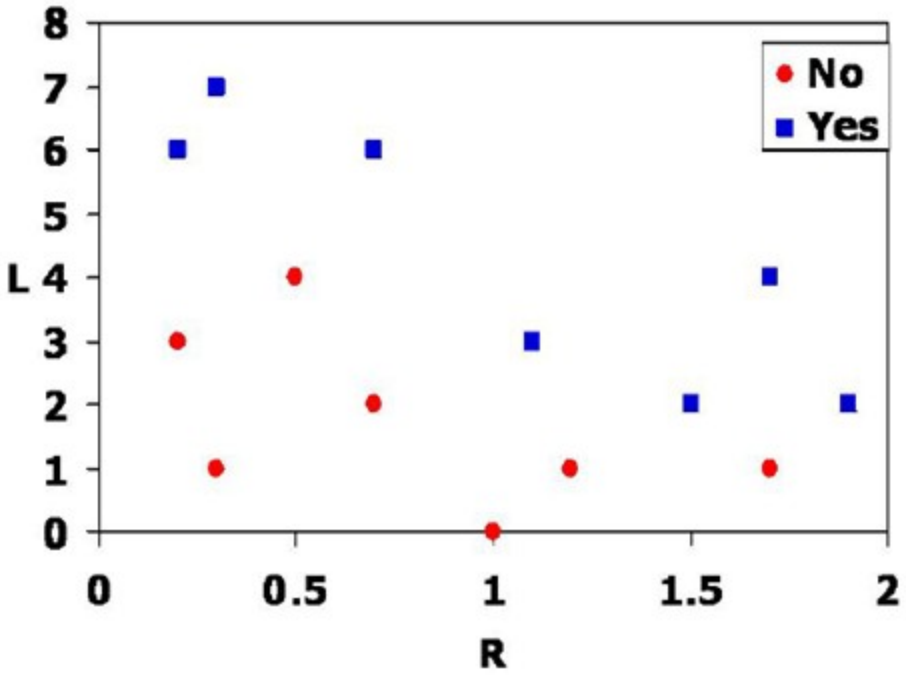
Knn pros and cons

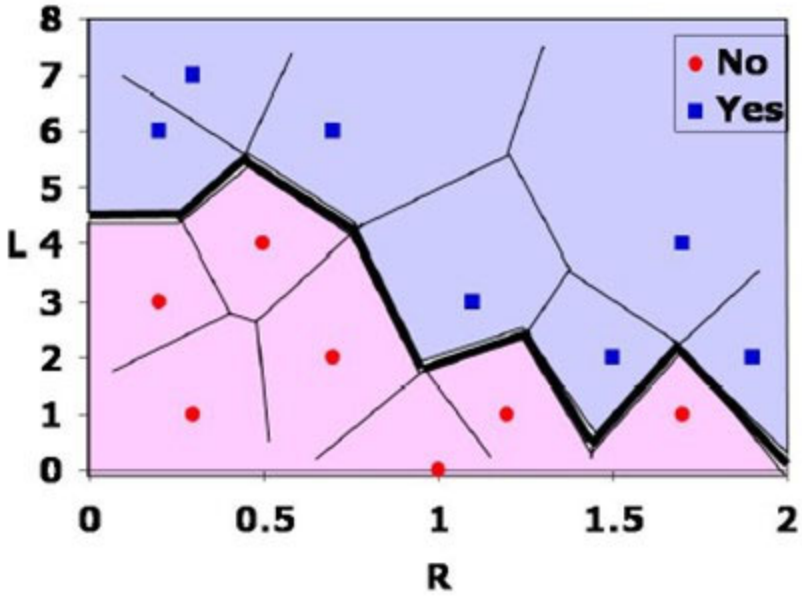
Decision boundaries



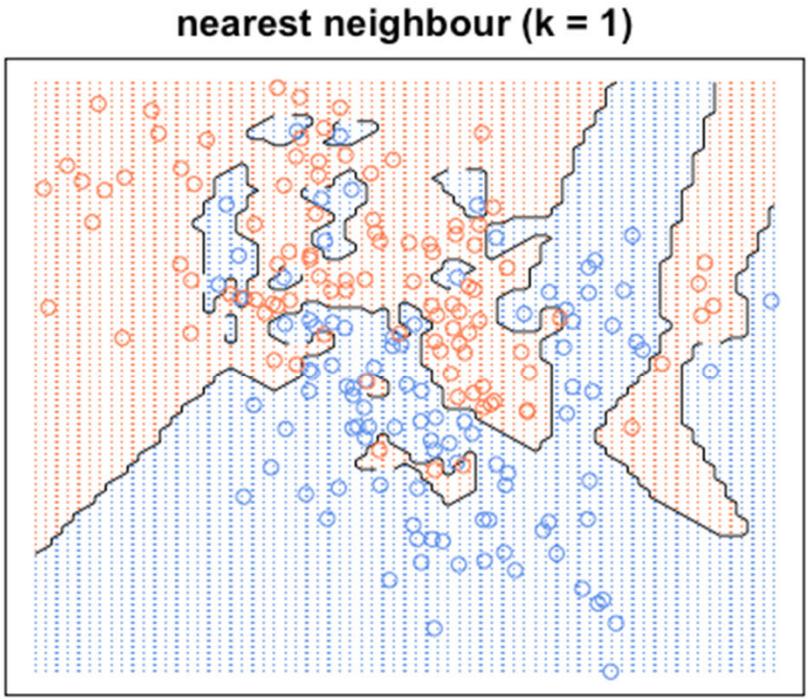
Decision boundaries

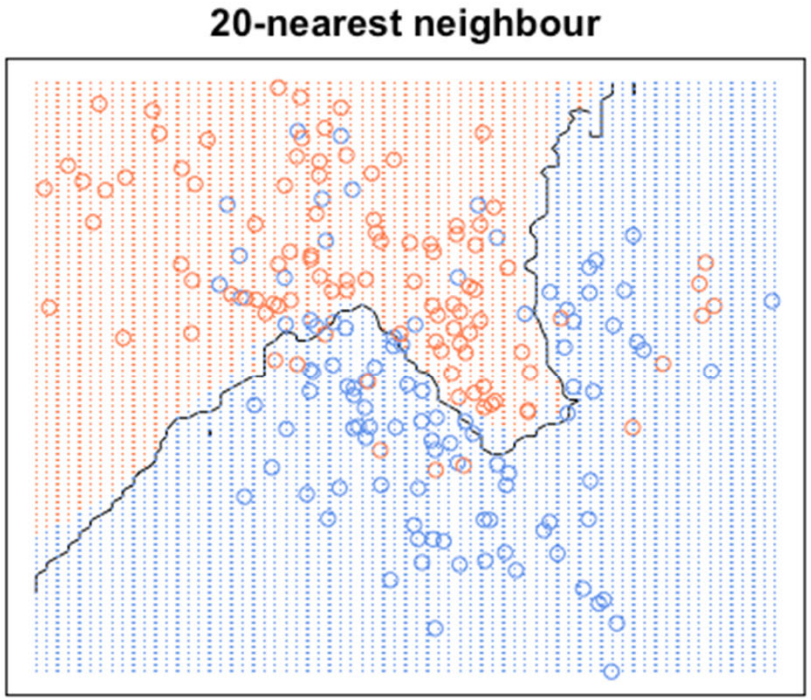
- Separate feature space into learned regions for each class



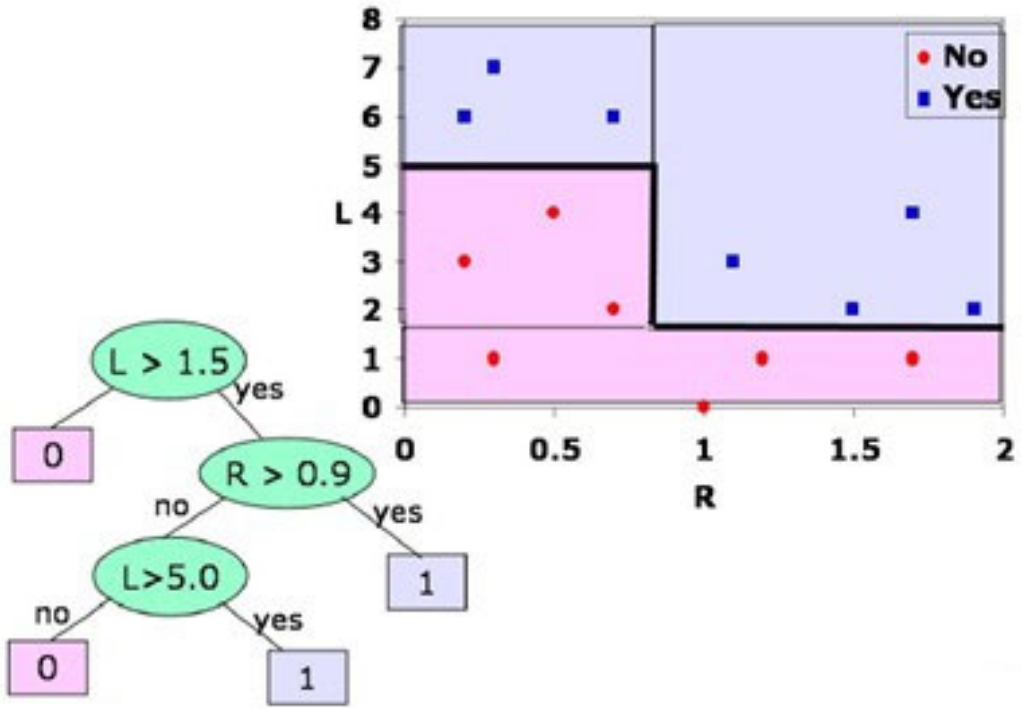


Knn decision boundaries

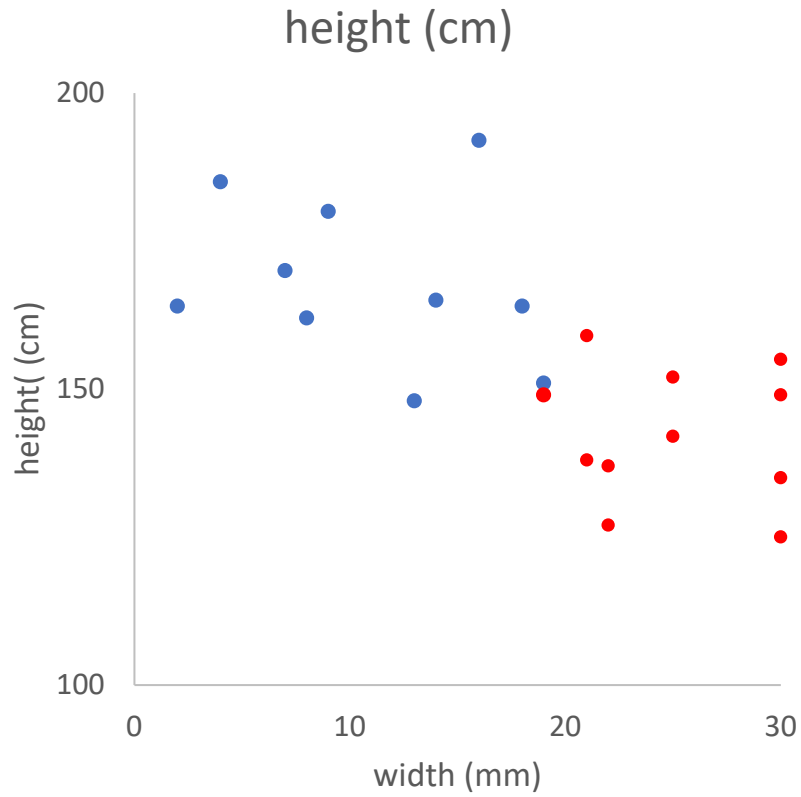


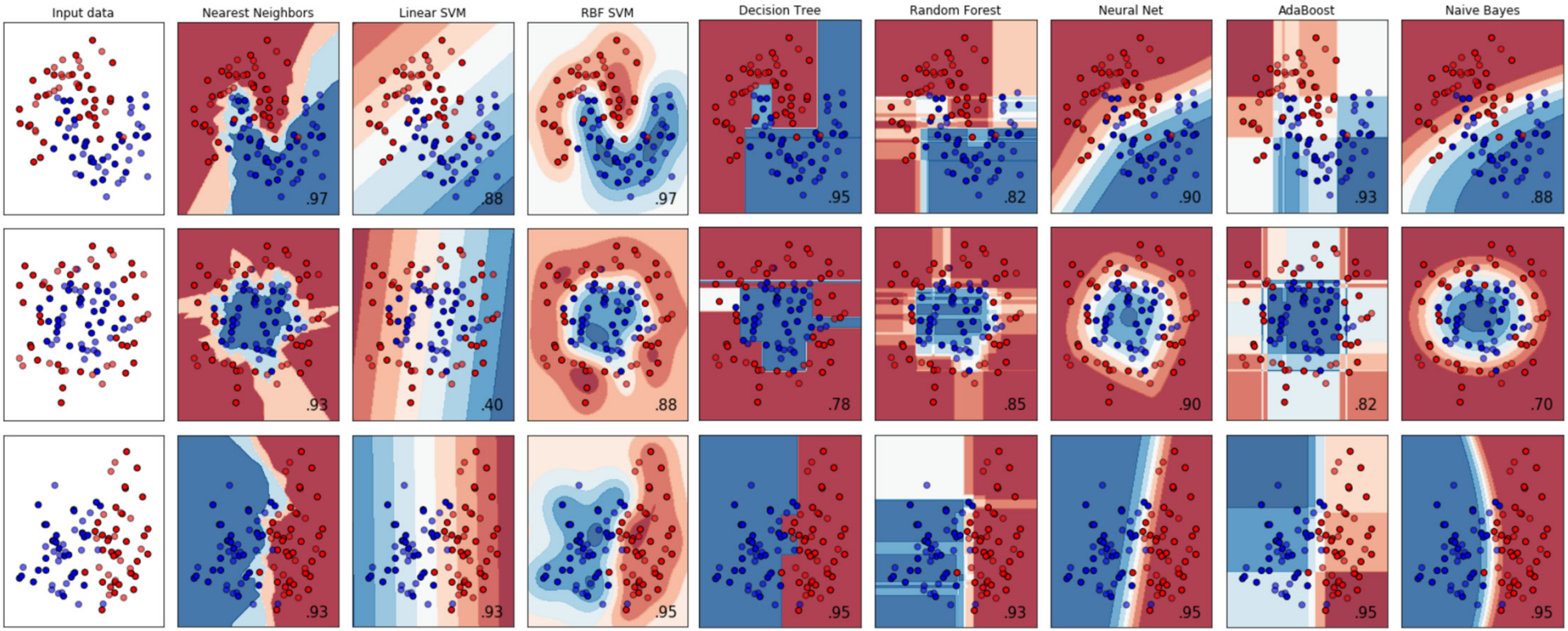


Decision tree decision boundaries



Decision tree decision boundary




Input data
Nearest Neighbors
Linear SVM
RBF SVM
Decision Tree
Random Forest
Neural Net
AdaBoost
Naive Bayes
.97
.88
.97
.95
.82
.90
.93
.88
.93
.40
.88
.78
.85
.90
.82
.70
.93
.93
.95
.95
.93
.95
.95
.95

Lets try this out