CS 171: Intro to ML and DM

Christian Shelton

UC Riverside

Slide Set 9: Nearest Neighbor II

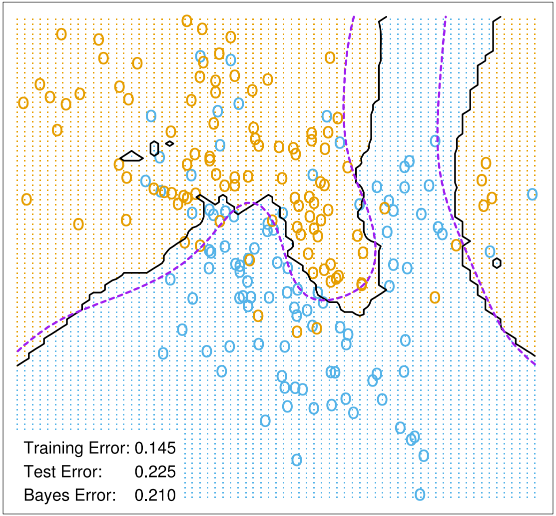
Slides from CS 171

- From UC Riverside
  - CS 171: Introduction to Machine Learning and Data Mining
  - Professor Christian Shelton
- DO NOT REDISTRIBUTE
  - These slides contain copyrighted material (used with permission) from
    - Elements of Statistical Learning (Hastie, et al.)
    - Pattern Recognition and Machine Learning (Bishop)
    - An Introduction to Machine Learning (Kubat)
    - Machine Learning: A Probabilistic Perspective (Murphy)
  - For use only by enrolled students in the course

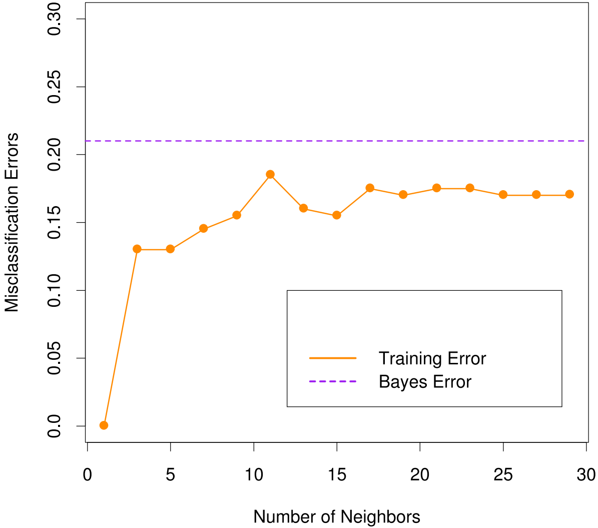
k-Nearest Neighbor Practical Considerations

- How to pick k
- How to pick distance measure (and scaling)
- How to reduce computational costs

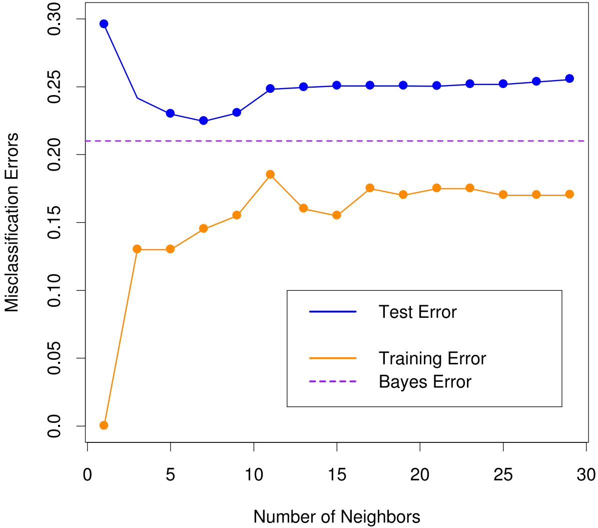
Selection of k



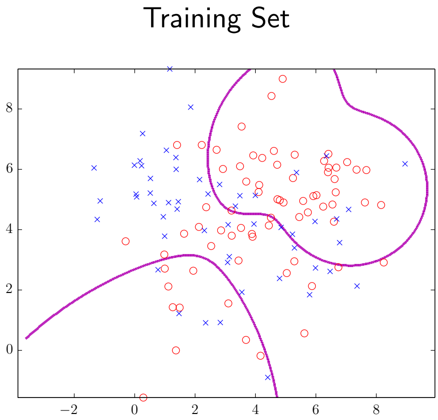
Selection of k

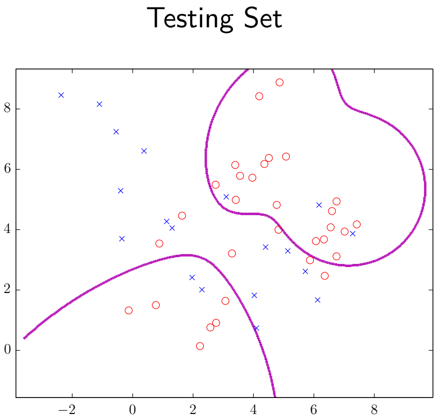


Selection of k



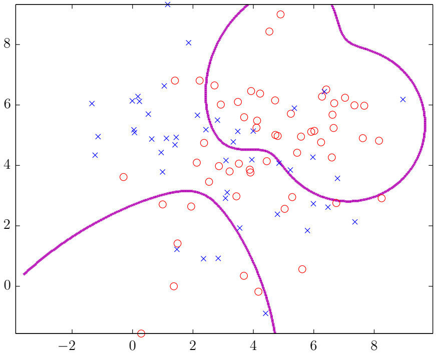
Cross Validation Example



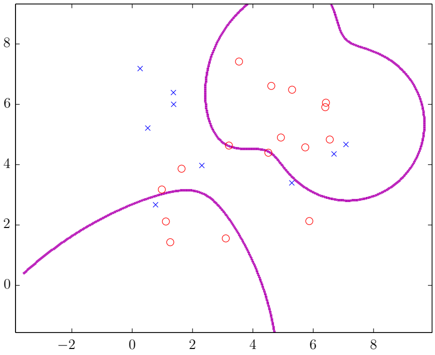


Cross Validation Example

Training Set

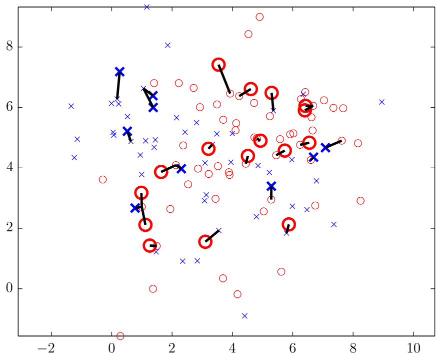


Validation Set

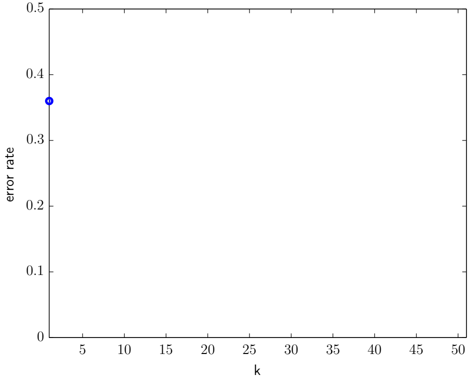


Cross Validation Example

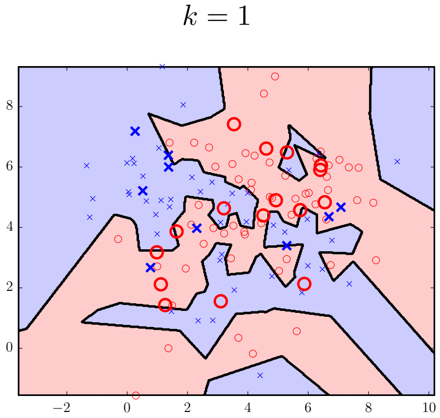
$k = 1$

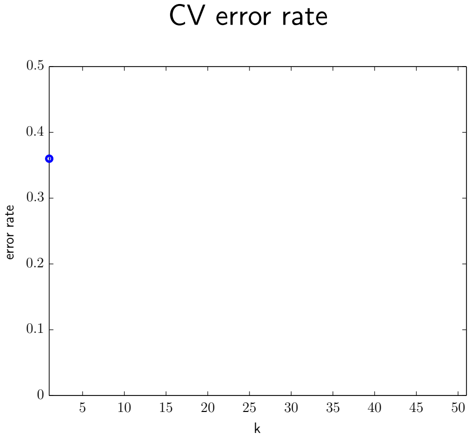


CV error rate



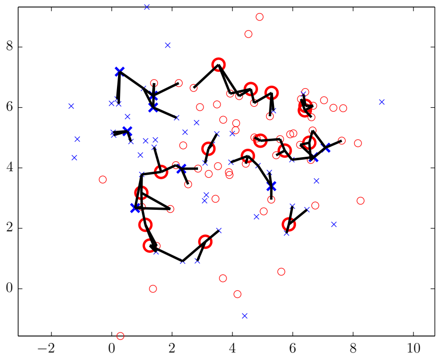
Cross Validation Example



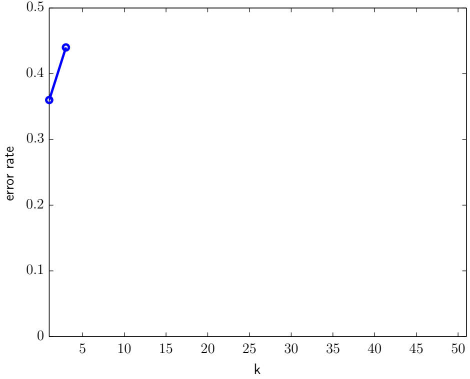


Cross Validation Example

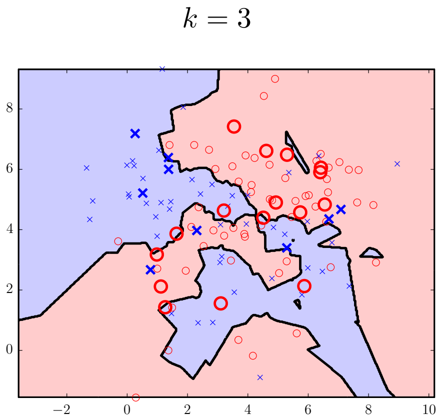
$k = 1$

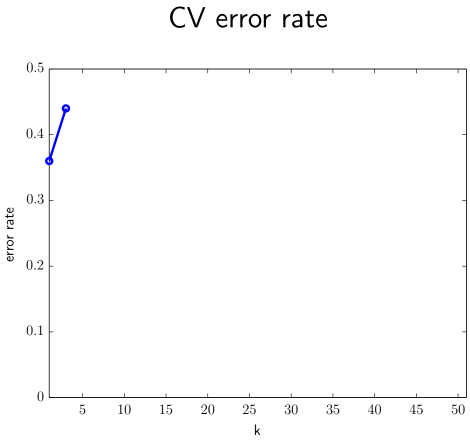


CV error rate



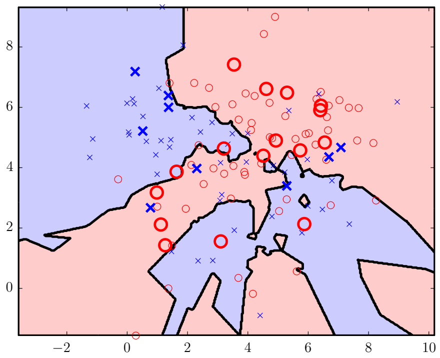
Cross Validation Example



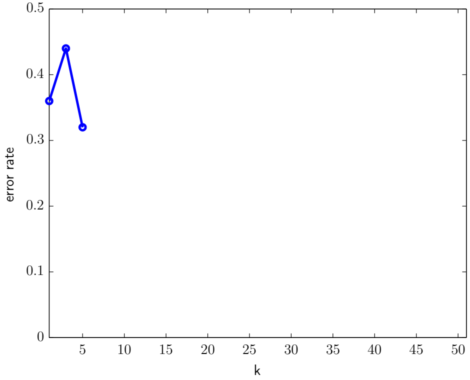


Cross Validation Example

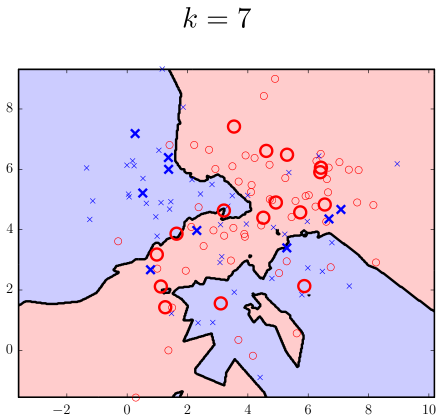
$k = 5$

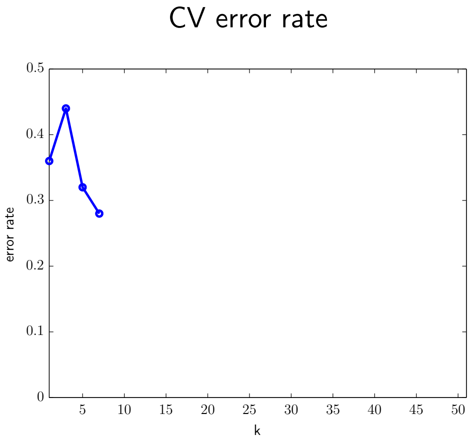


CV error rate

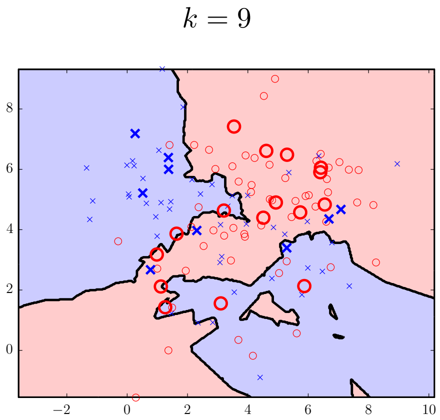


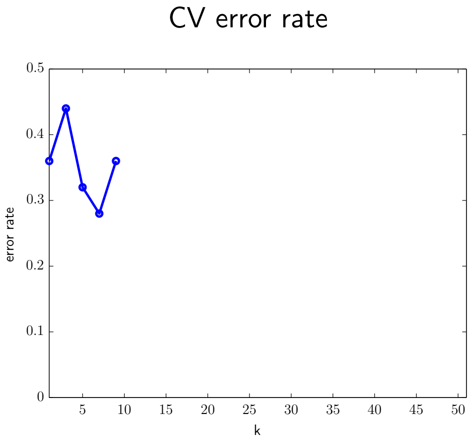
Cross Validation Example





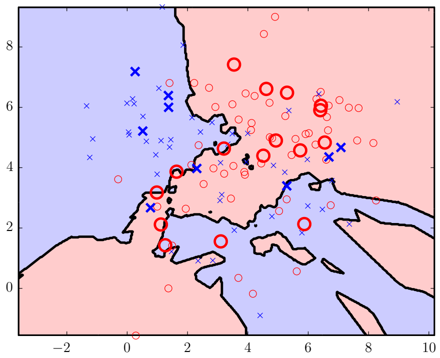
Cross Validation Example



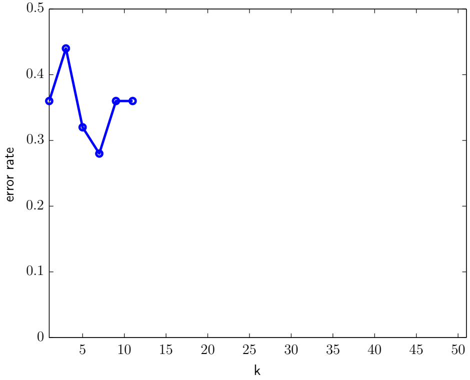


Cross Validation Example

$k = 11$

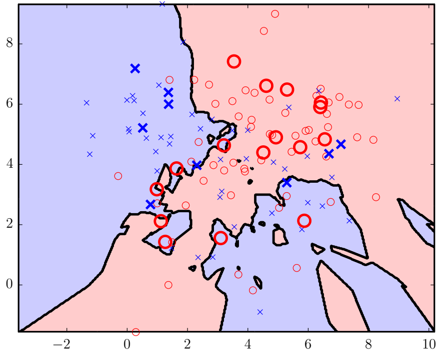


CV error rate

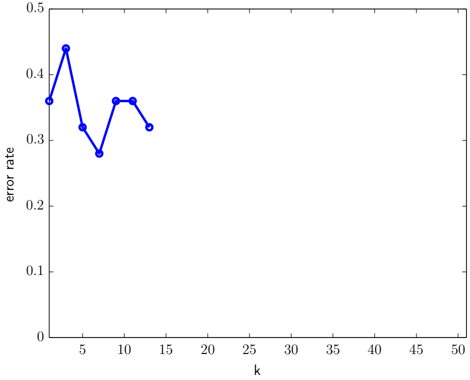


Cross Validation Example

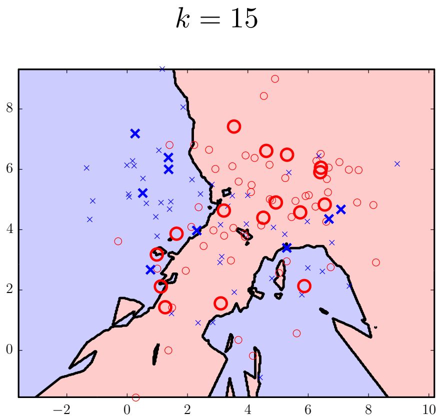
$k = 13$

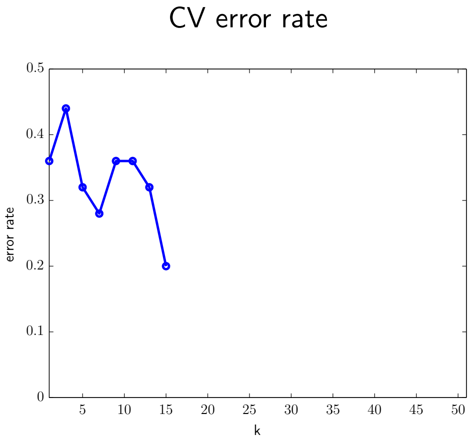


CV error rate



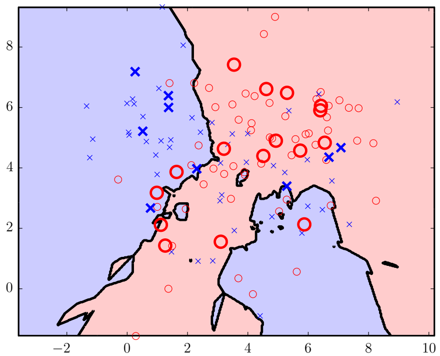
Cross Validation Example



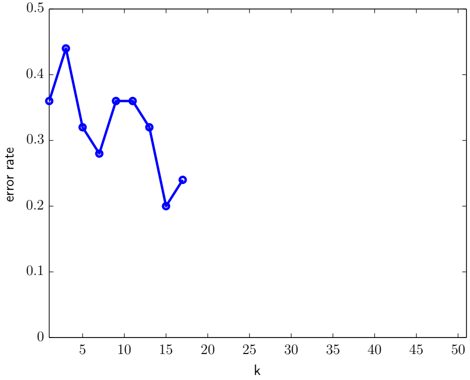


Cross Validation Example

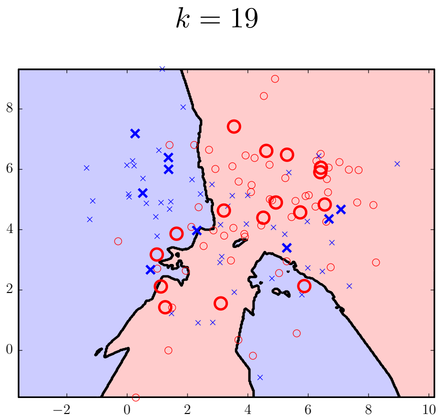
$k = 17$

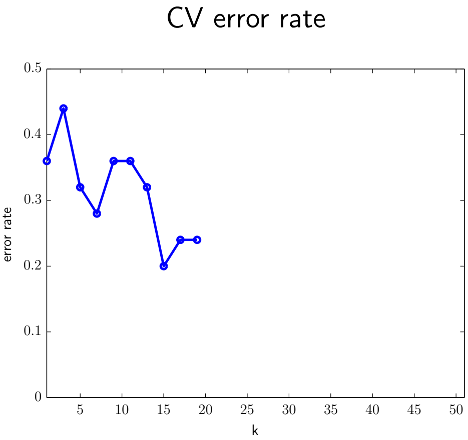


CV error rate



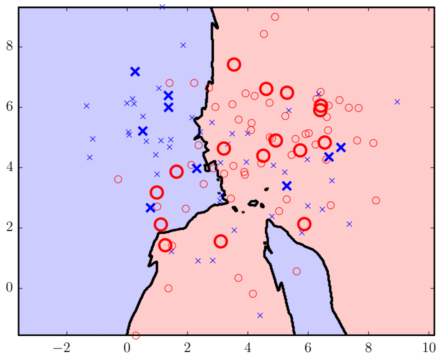
Cross Validation Example



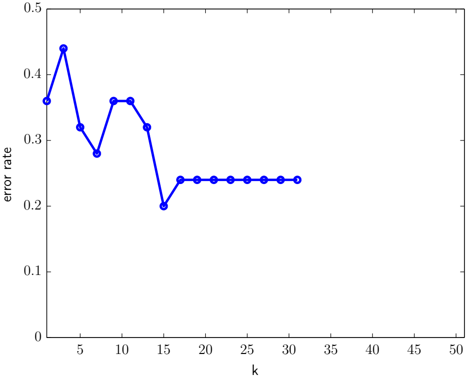


Cross Validation Example

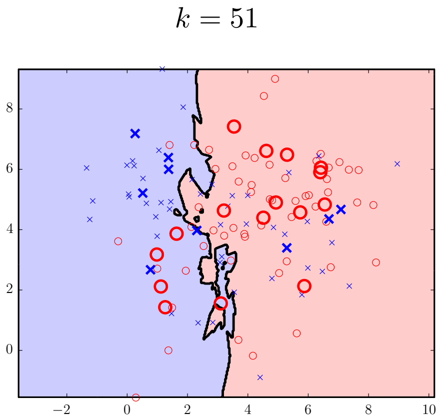
$k = 31$

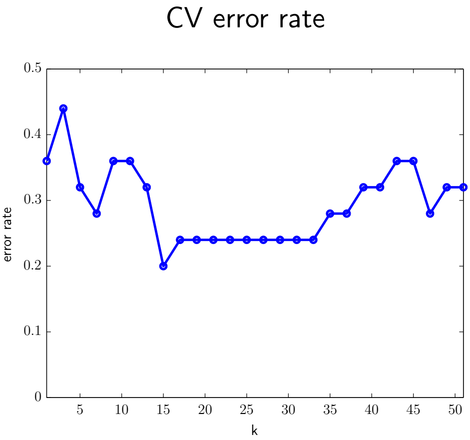


CV error rate



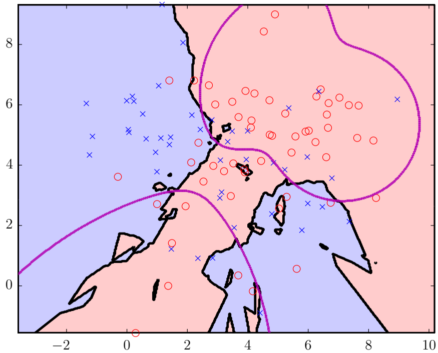
Cross Validation Example



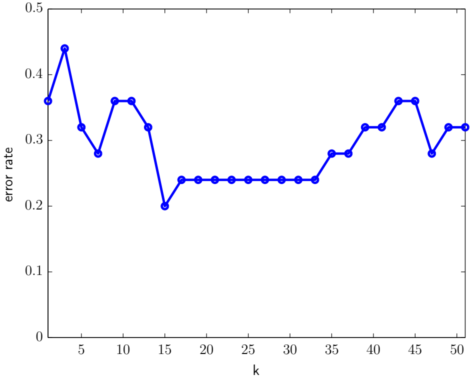


Cross Validation Example

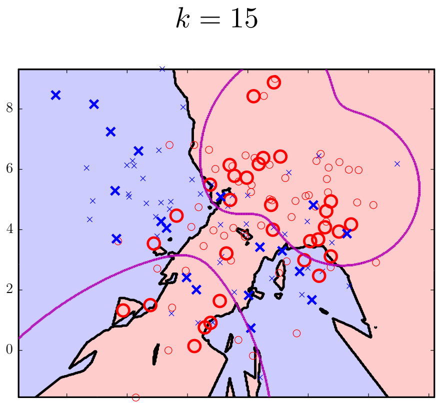
$k = 15$

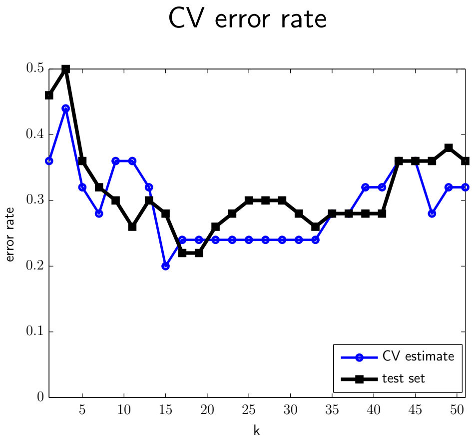


CV error rate

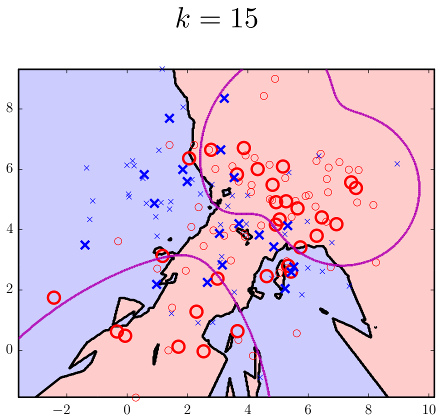


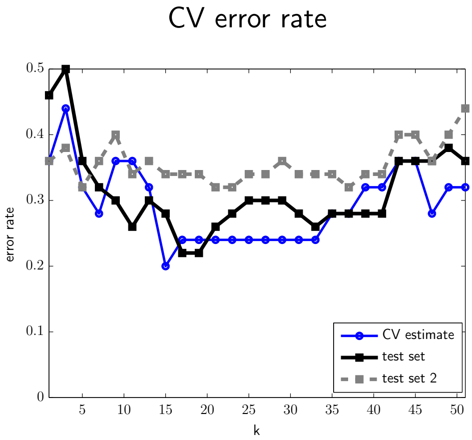
Cross Validation Example



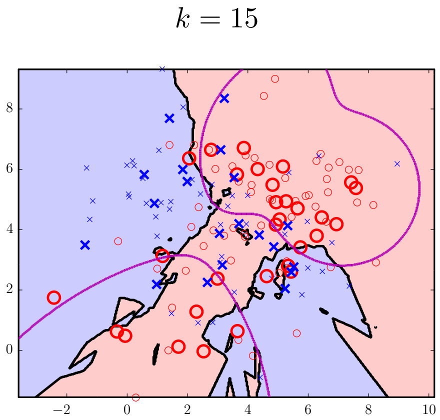


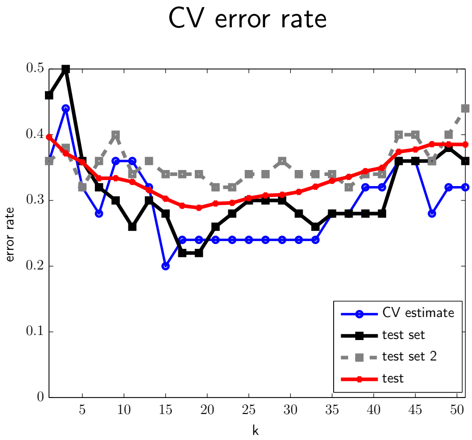
Cross Validation Example



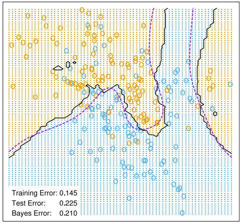


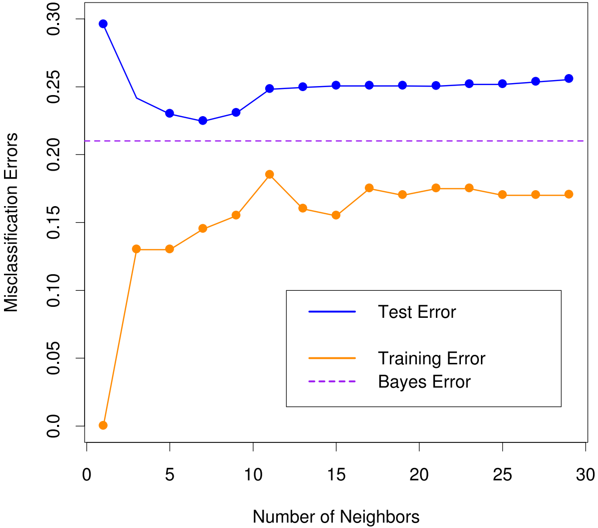
Cross Validation Example



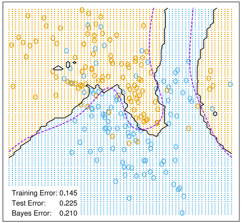


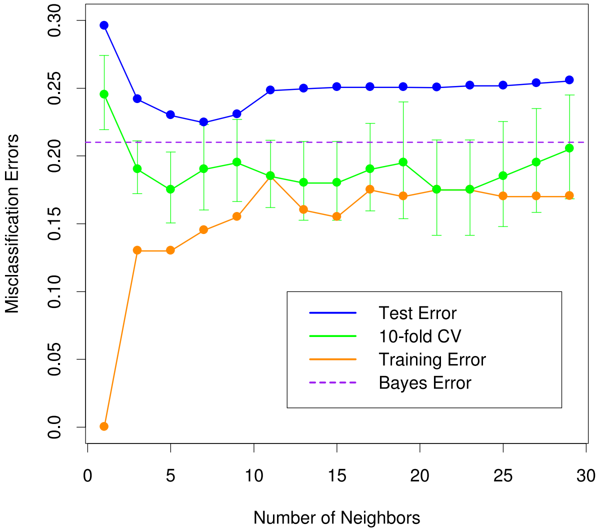
Selection of k





Selection of k





Selection of Distance Metric

- Use cross validation

Selection of Distance Metric

- Use cross validation
- Fine unless you have lots of distance metrics to try (overfitting to the validation set!)
- If distance metric also includes axis scaling, need something else

Attribute Scaling

- Distance metric learning algorithms (beyond the scope of this course)

Attribute Scaling

- Distance metric learning algorithms (beyond the scope of this course)
- "Normalize" the attributes
  - Scale and translate so that the min is 0 and the max is 1:

$$\mathrm{min}_j = \min_i x_{i,j}$$

$$\mathrm{max}_j = \max_i x_{i,j}$$

$$x_{i,j} \leftarrow \frac{x_{i,j} - \mathrm{min}_j}{\mathrm{max}_j - \mathrm{min}_j}$$

Attribute Scaling

- Distance metric learning algorithms (beyond the scope of this course)
- "Normalize" the attributes
  - Scale and translate so that the min is 0 and the max is 1:

$$\mathsf{min}_j = \min_i x_{i,j}$$

$$\mathsf{max}_j = \max_i x_{i,j}$$

$$x_{i,j} \leftarrow \frac{x_{i,j} - \mathsf{min}_j}{\mathsf{max}_j - \mathsf{min}_j}$$

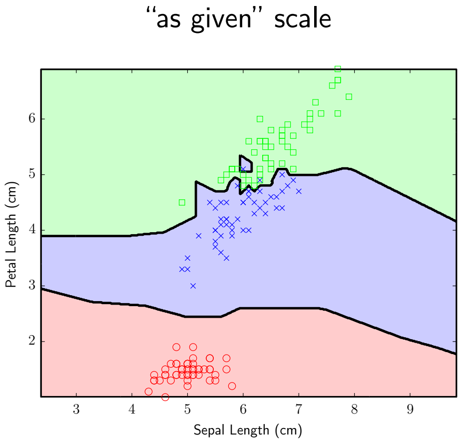
  - Scale and translate so that the mean is 0 and the standard deviation is 1 (z-normalization and many other names)

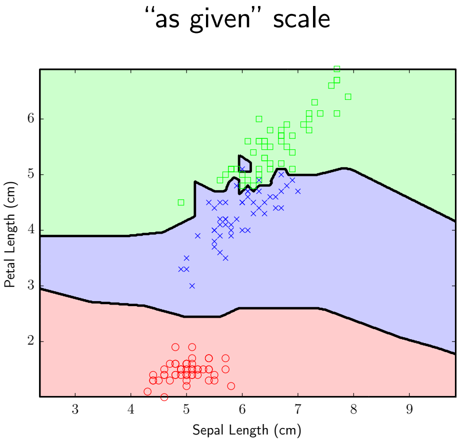
$$\mathsf{mean}_j = \frac{1}{m}\sum_i x_{i,j}$$

$$\mathsf{std}_j = \sqrt{\frac{1}{m}\sum_i (x_{i,j} - \mathsf{mean}_j)^2}$$

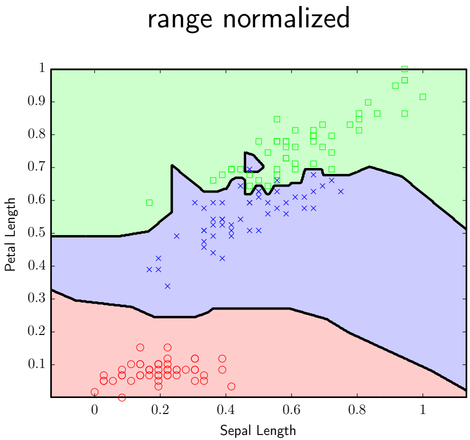
$$x_{i,j} \leftarrow \frac{x_{i,j} - \mathsf{mean}_j}{\mathsf{std}_j}$$

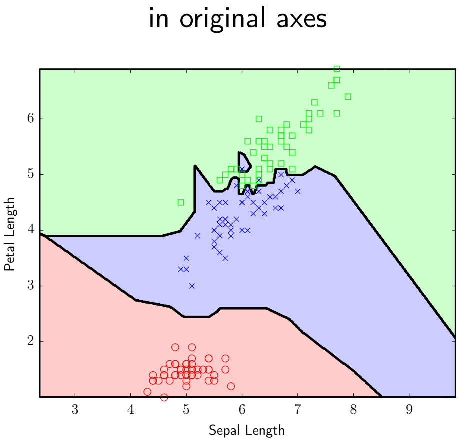
Iris Scaling



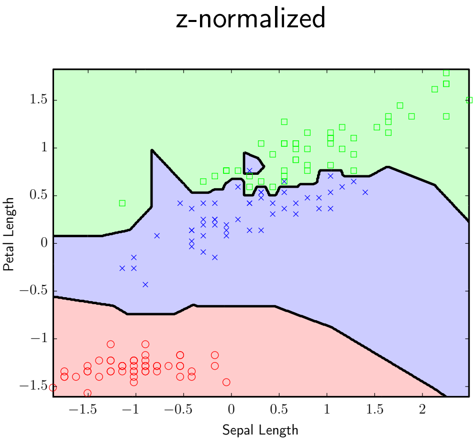


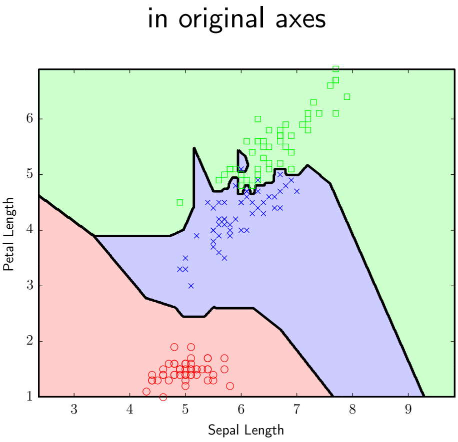
Iris Scaling





Iris Scaling





Example Problem

Problem specification: to classify land use from satelite imagery.

Example Problem

Problem specification: to classify land use from satelite imagery.

- Each pixel corresponds to an area on the Earth.
- Observations are the intensity from four different spectral bands (two visible spectrum, two infrared)
- 7 different classes: red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil, very damp gray soil

Example Problem

Problem specification: to classify land use from satelite imagery.

- Each pixel corresponds to an area on the Earth.
- Observations are the intensity from four different spectral bands (two visible spectrum, two infrared)
- 7 different classes: red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil, very damp gray soil

Attributes: the four values from the pixel in question,

Example Problem

Problem specification: to classify land use from satelite imagery.

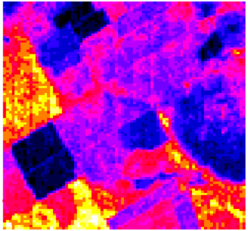
- Each pixel corresponds to an area on the Earth.
- Observations are the intensity from four different spectral bands (two visible spectrum, two infrared)
- 7 different classes: red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil, very damp gray soil

Attributes: the four values from the pixel in question,
plus the four values from each of the neighboring eight pixels.
(total 36 attributes)

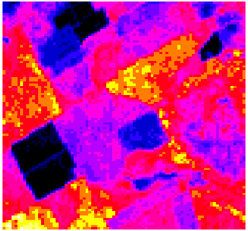
Example Problem

Using $k = 5$, performance is better than any other tried method (error rate of 9.5%).

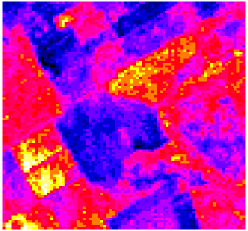
Spectral Band 1



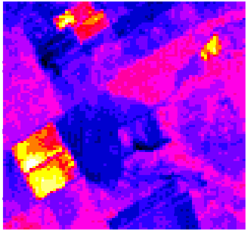
Spectral Band 2



Spectral Band 3



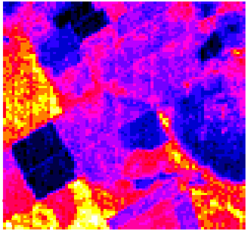
Spectral Band 4



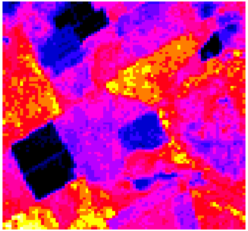
Example Problem

Using $k = 5$, performance is better than any other tried method (error rate of 9.5%).

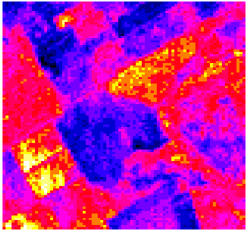
Spectral Band 1



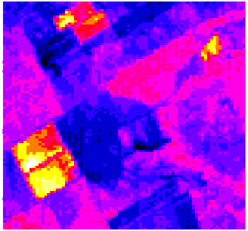
Spectral Band 2



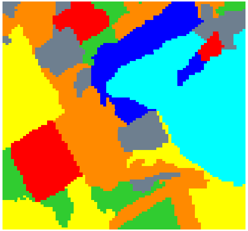
Spectral Band 3



Spectral Band 4

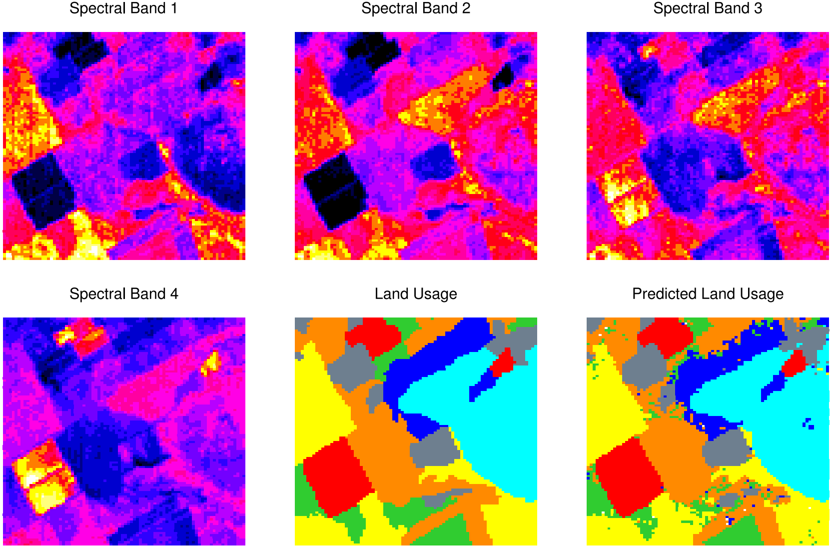


Land Usage



Example Problem

Using $k = 5$, performance is better than any other tried method (error rate of 9.5%).



Computational Concerns

The time to test a single point is $O(m)$ (linear in the size of the training data).

Computational Concerns

The time to test a single point is $O(m)$ (linear in the size of the training data). (or is it?)

Computational Concerns

The time to test a single point is $O(m)$ (linear in the size of the training data). (or is it?)
This is a problem for large datasets.

Computational Concerns

The time to test a single point is $O(m)$ (linear in the size of the training data). (or is it?)

This is a problem for large datasets.

We can try to prune the dataset.

There are many ways...

Condensing

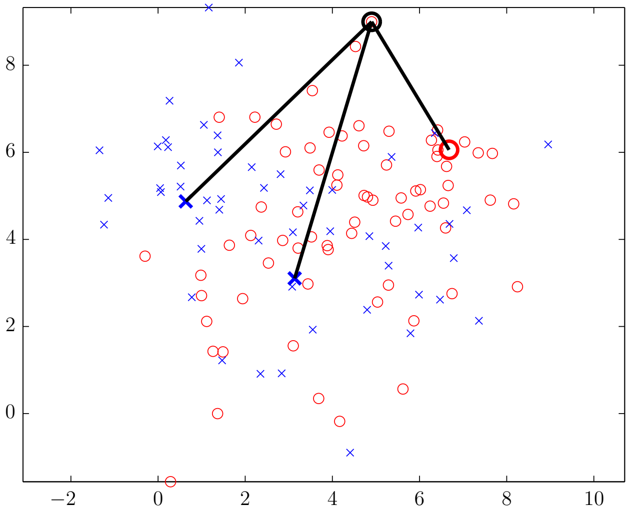
1. Add a single random point to the current point set
2. In random order:
    1. Take a point and classify it according to the current point set
    2. If it is incorrectly classified, add it to the current point set

Condensing

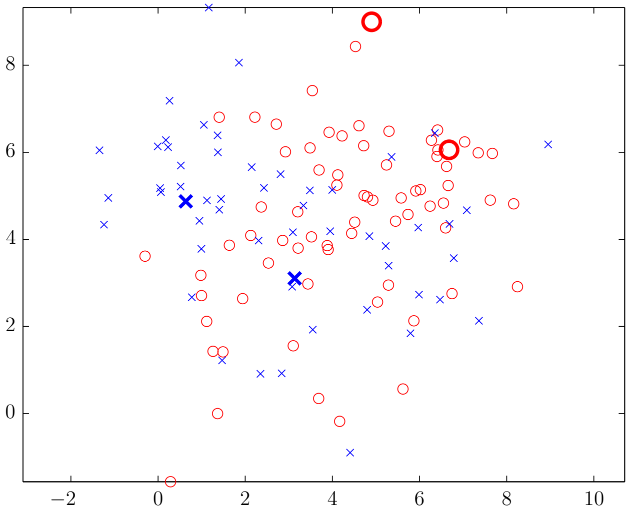
1. Add a single random point to the current point set
2. In random order:
    1. Take a point and classify it according to the current point set
    2. If it is incorrectly classified, add it to the current point set

May need to be repeated to find consistent set

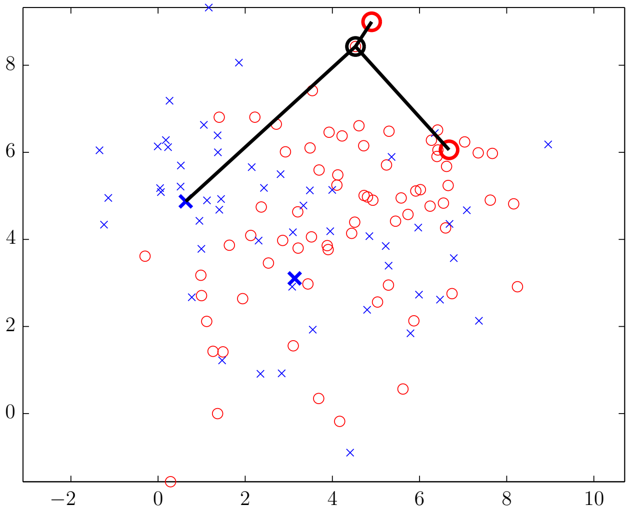
Condensing



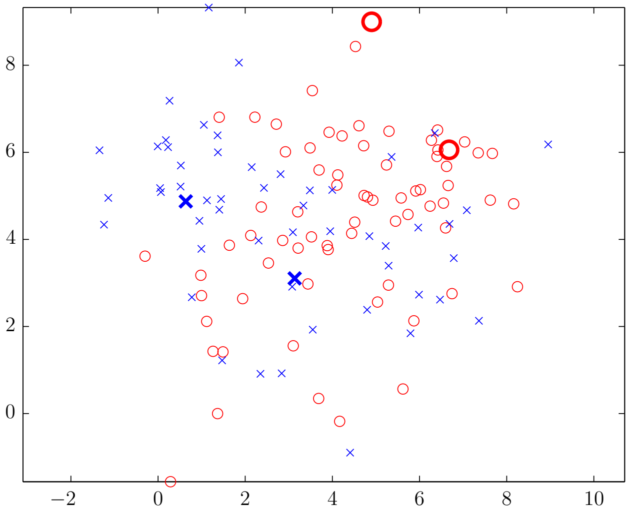
Condensing



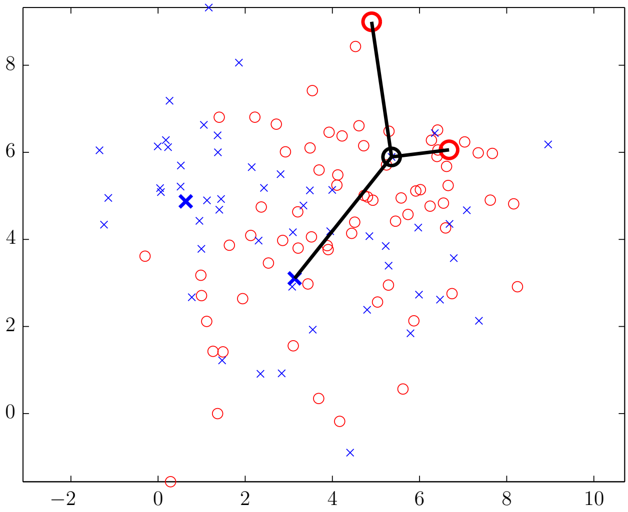
Condensing



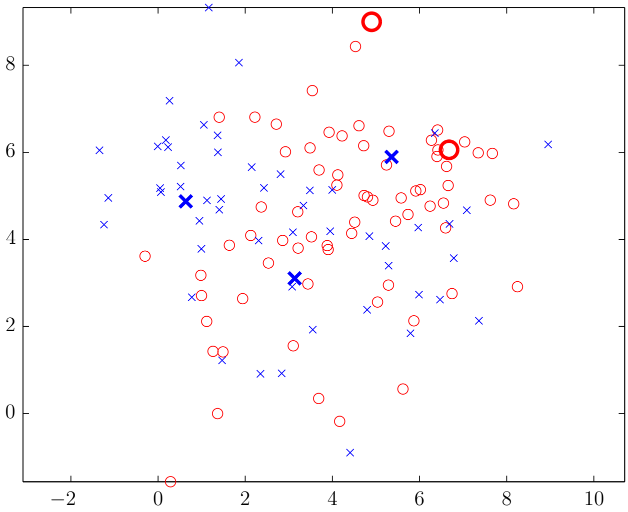
Condensing



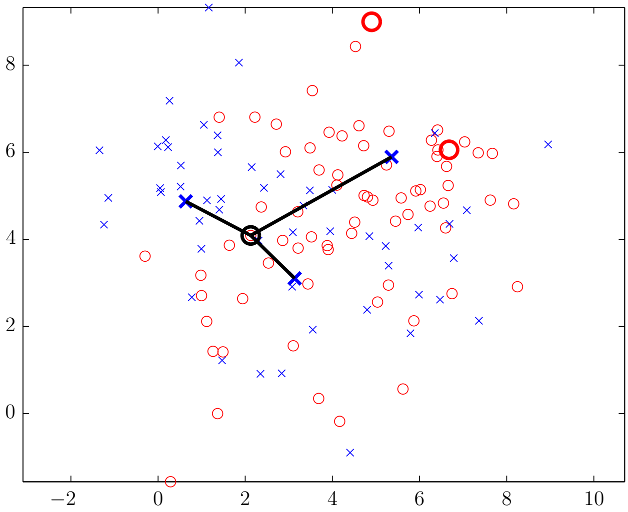
Condensing



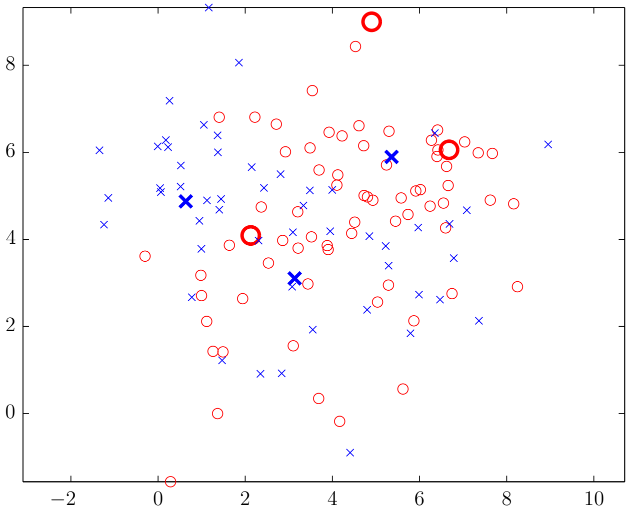
Condensing



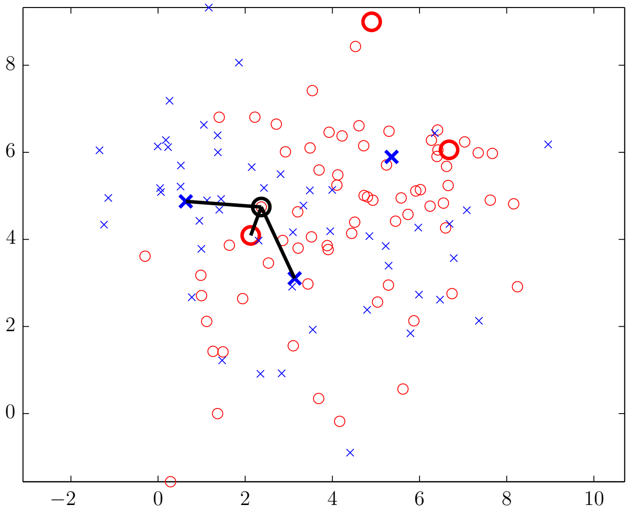
Condensing



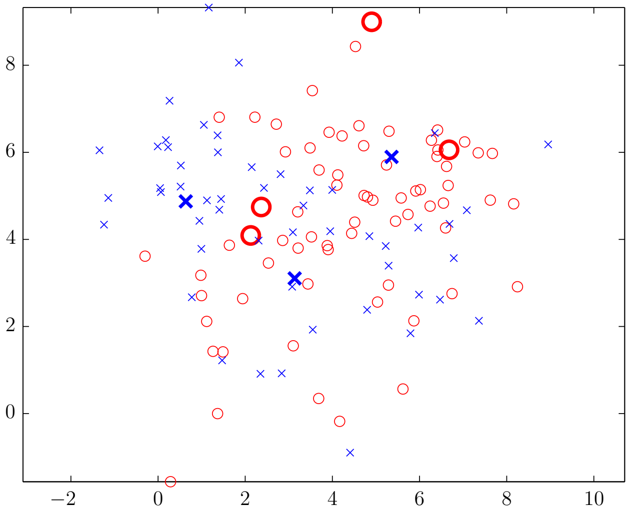
Condensing



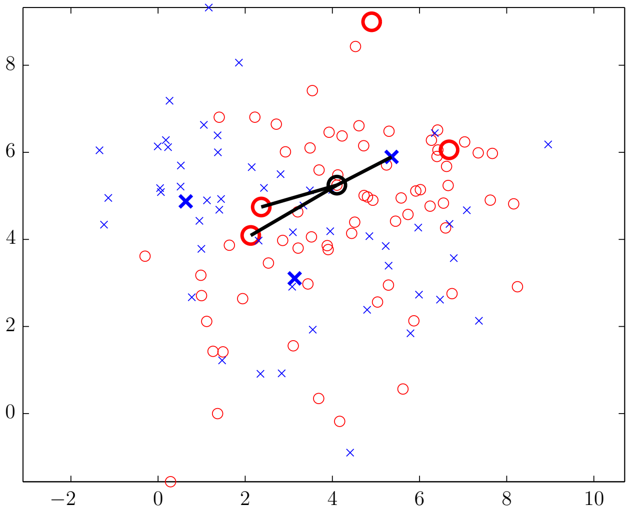
Condensing



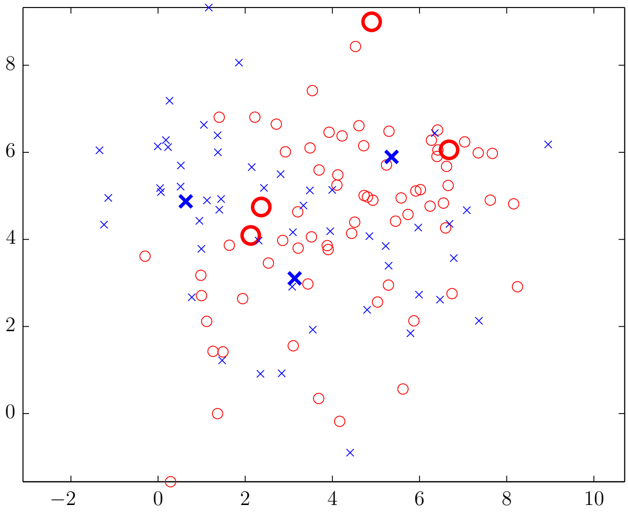
Condensing



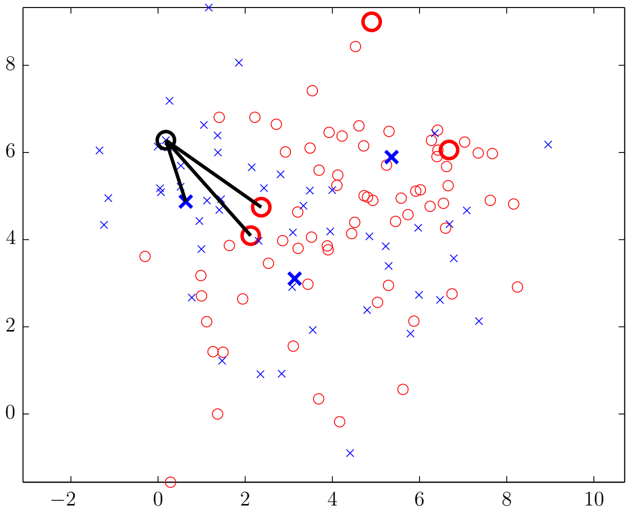
Condensing



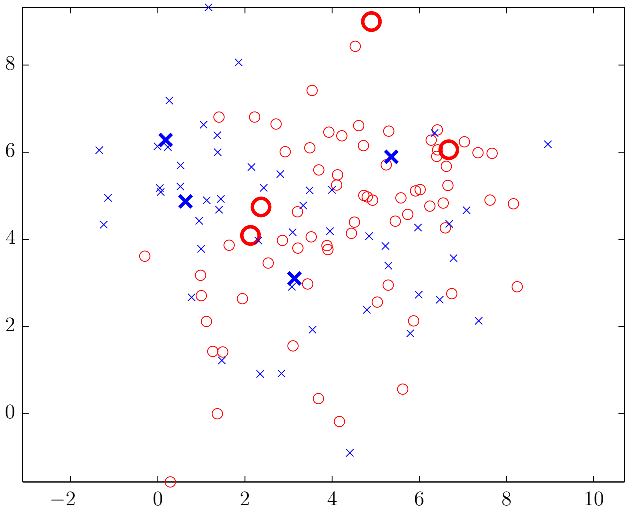
Condensing



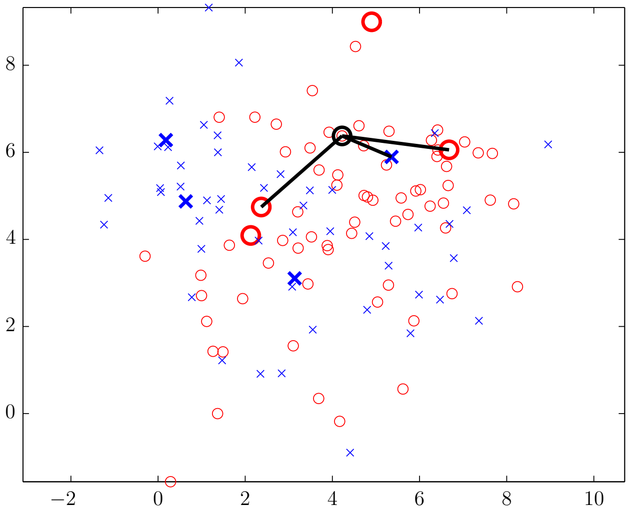
Condensing



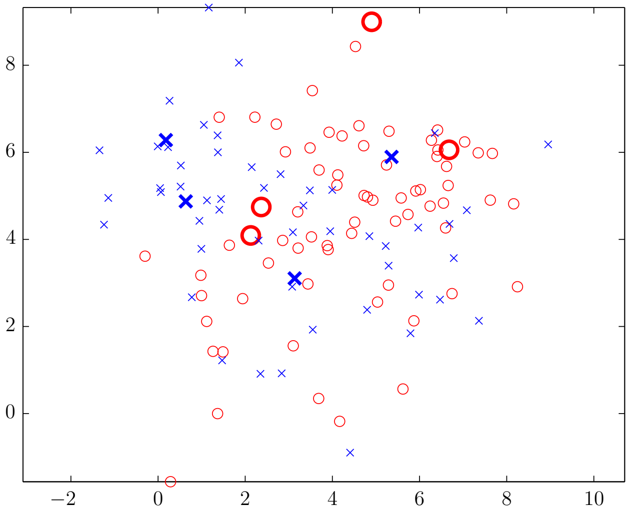
Condensing



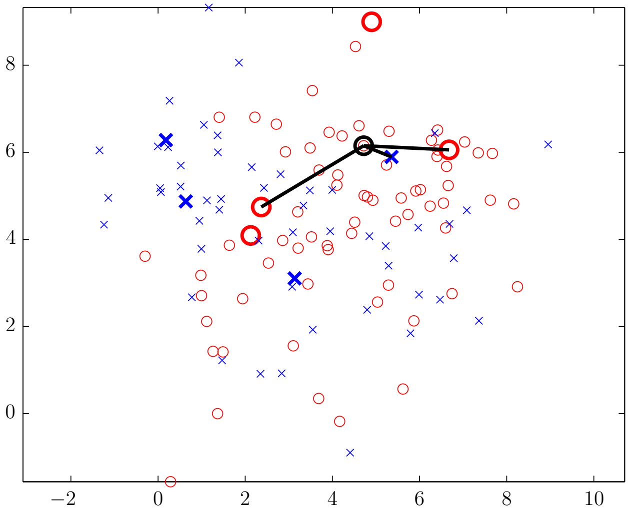
Condensing



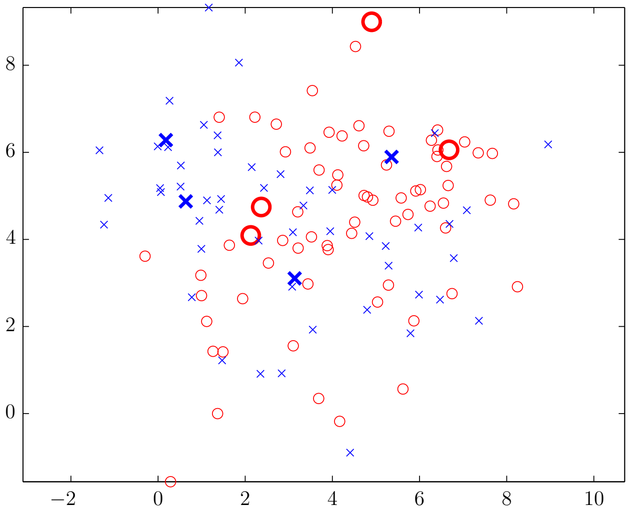
Condensing



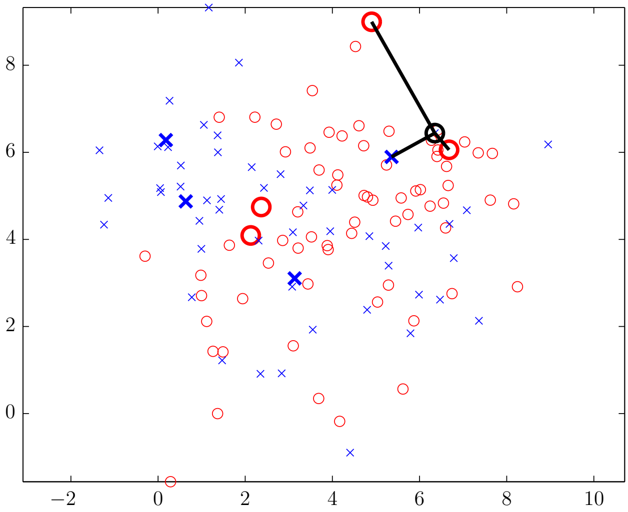
Condensing



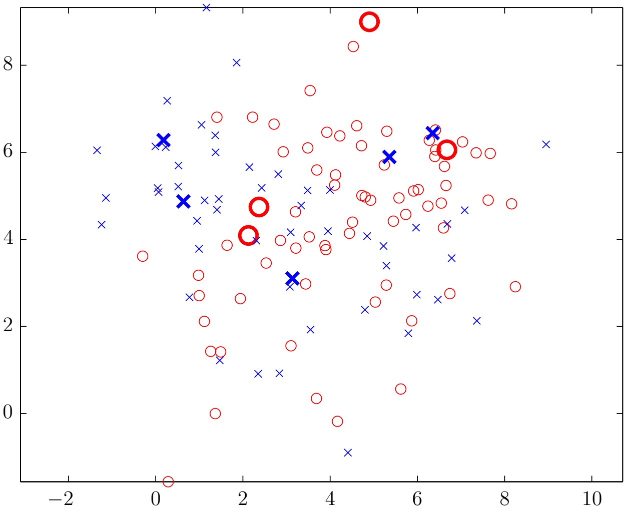
Condensing



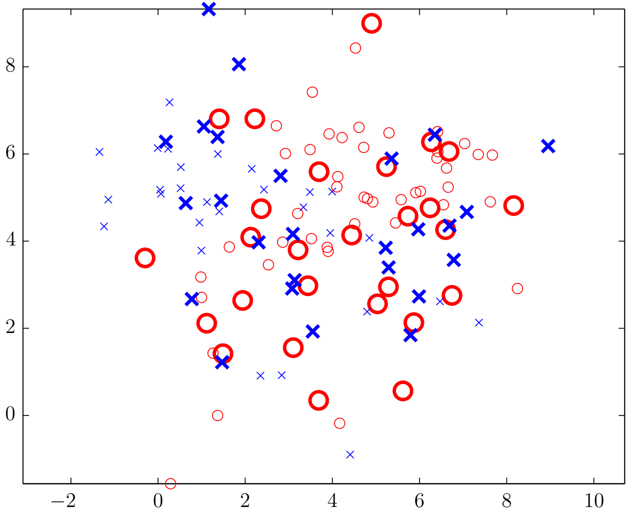
Condensing



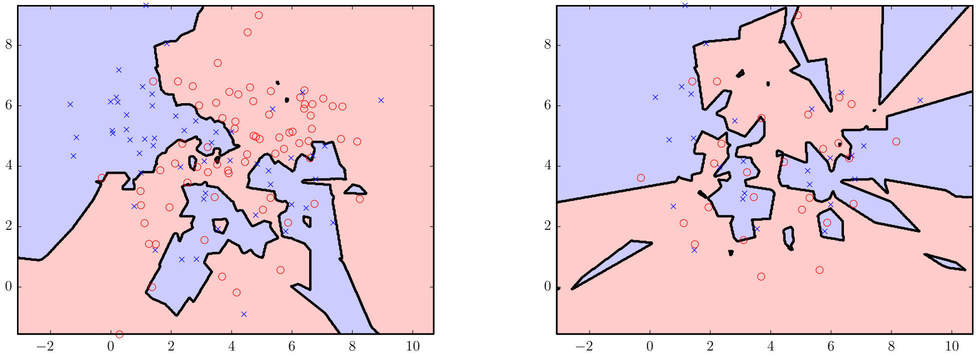
Condensing



Condensing



Condensing



k-Nearest Neighbor

When is it good?

- Low-dimensional space
- Lots of data
- Lots of computational power
- Highly irregular decision surface