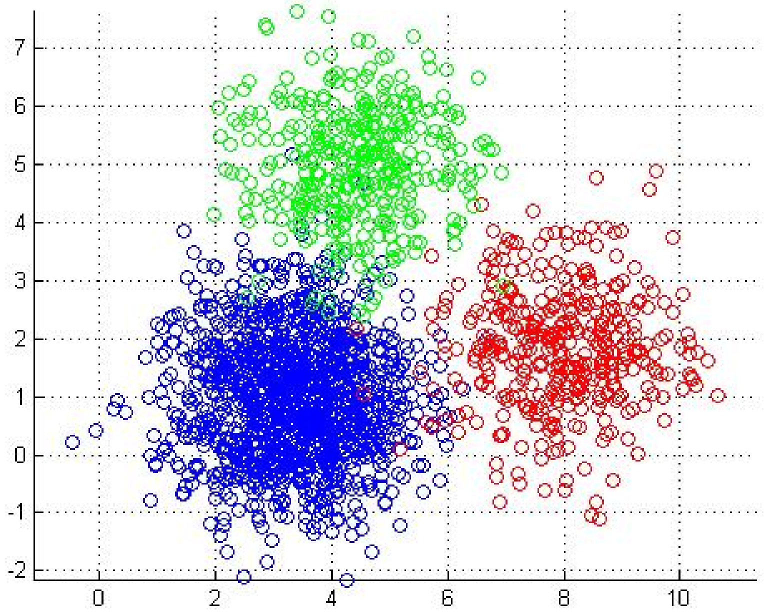
k-means



mathworks.com

Note that this is not related to kNN.

Explain:

- Choose k
- Choose k centroids
- Compute distances and assign each point to closest centroid
- Now re-compute centroids based on classes
- Now re-assign based on closest
- Converges because distances get smaller

Parameter sweep is a possibility for choosing k.

Assumes roughly spherical clusters.

Assignment:

- Pick k random centroids from sample (*Forgy* method).
- Assign to k random clusters (*random partitions* method).

Uses

- Vector quantization (picking hopefully prototypical samples)

Examples

Code time

Discussion

When should we use k-means vs logistic regression?

1 SVM

SVM

Support Vector Machines

- Goal: optimal separating hyperplane
- aka: Large Margin Classifier

Discussion:

Consider the example of dots at [[0, 1], [1, 0]].

Strategies

At the beginning, one tends to do this:

- Transform data for SVM solver
- Randomly try a few kernels and parameters
- Test

A better strategy:

- Transform
- Scale data
- Consider linear, Gaussian, or RBF kernels
- Use cross validation to find the best C and γ
- Test

Grid search often works well for C and γ. Try exponentially increasing C and decreasing γ. E.g., $2^{-5}, 2^{-3}, \ldots, 2^{15}$. Categories usually work better as binary fields than as enums.

If there are lot of features, linear often works well.

n features, m training examples.

When n (e.g., 10^4) is much larger m (e.g., 10 to 10^3), then linear regression or SVM with linear kernel tends to work well.

When n is small ($n < 10^3$) and m medium ($m < 10^4$), then Gaussian kernel often works well.

When m is large ($n < 10^3$ and $m > 5 \cdot 10^4$), add features and then use linear regression or SVM with a linear kernel.

2 Break

Break

Questions?

`purple.com/talk-feedback`