Naïve Bayes

Likelihood

Class Prior Probability

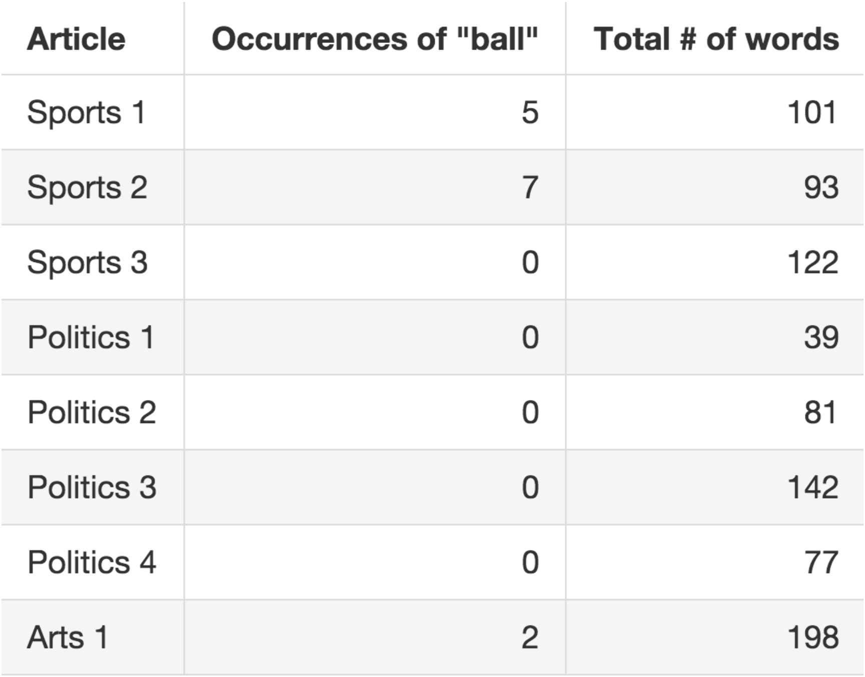
$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = \underbrace{P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c)}_{\text{how Naïve!}} \times P(c)$$

Naïve Bayes



Article	Occurrences of "ball"	Total # of words
Sports 1	5	101
Sports 2	7	93
Sports 3	0	122
Politics 1	0	39
Politics 2	0	81
Politics 3	0	142
Politics 4	0	77
Arts 1	2	198

$$P(\text{“ball”}|\text{sports}) = \frac{5+7+0}{101+93+122} = \frac{12}{316} = 0.038$$

$$P(\text{“ball”}|\text{politics}) = \frac{0+0+0+0}{39+81+142+77} = \frac{0}{339} = 0.0$$

$$P(\text{“ball”}|\text{arts}) = \frac{2}{198} = 0.010$$

Naïve Bayes

Which category for very short article “the giants beat the nationals”?

$$
\begin{aligned}
P(\text{sports}|X) = & P(\text{sports}) \\
& \times P(\text{“the”}|\text{sports}) \\
& \times P(\text{“giants”}|\text{sports}) \\
& \times P(\text{“beat”}|\text{sports}) \\
& \times P(\text{“the”}|\text{sports}) \\
& \times P(\text{“nationals”}|\text{sports})
\end{aligned}
$$

Don’t forget LaPlace smoothing….

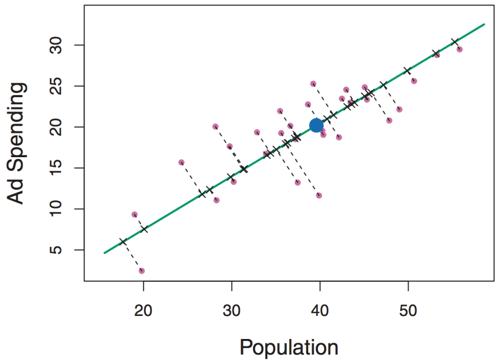
$$
P(x|c) = \frac{(\text{\# of times } x \text{ appears in articles of class } c) + \alpha}{(\text{total \# of words in articles of class } c) + \alpha \cdot (\text{\# of words in corpus})}
$$

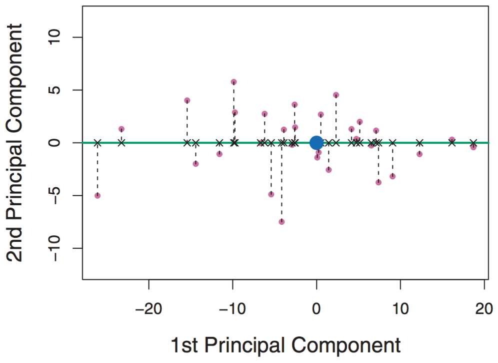
Unsupervised Learning

Two most common and contrasting unsupervised techniques

PCA

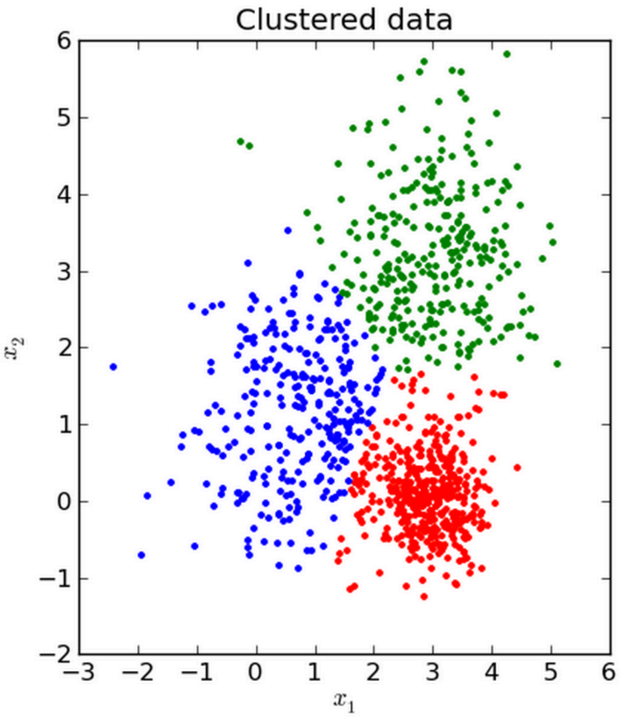
Low-dim representation of data that explains good fraction of variance





Clustering

Find homogenous subgroups among data

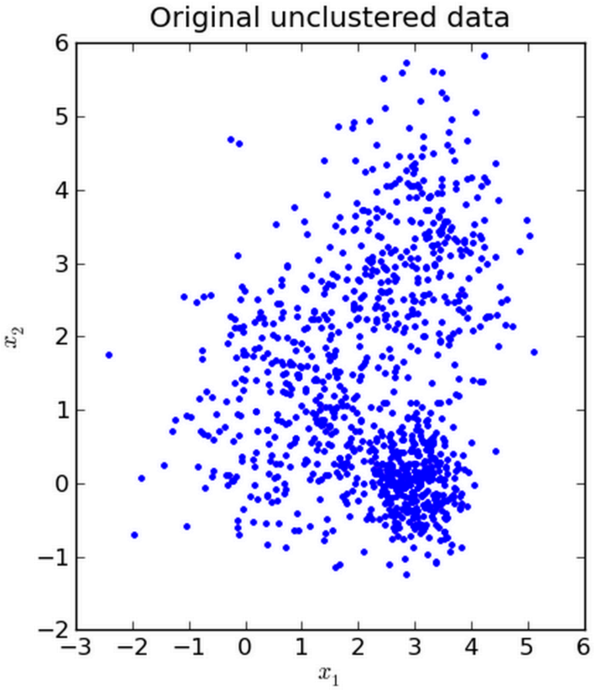


K-means

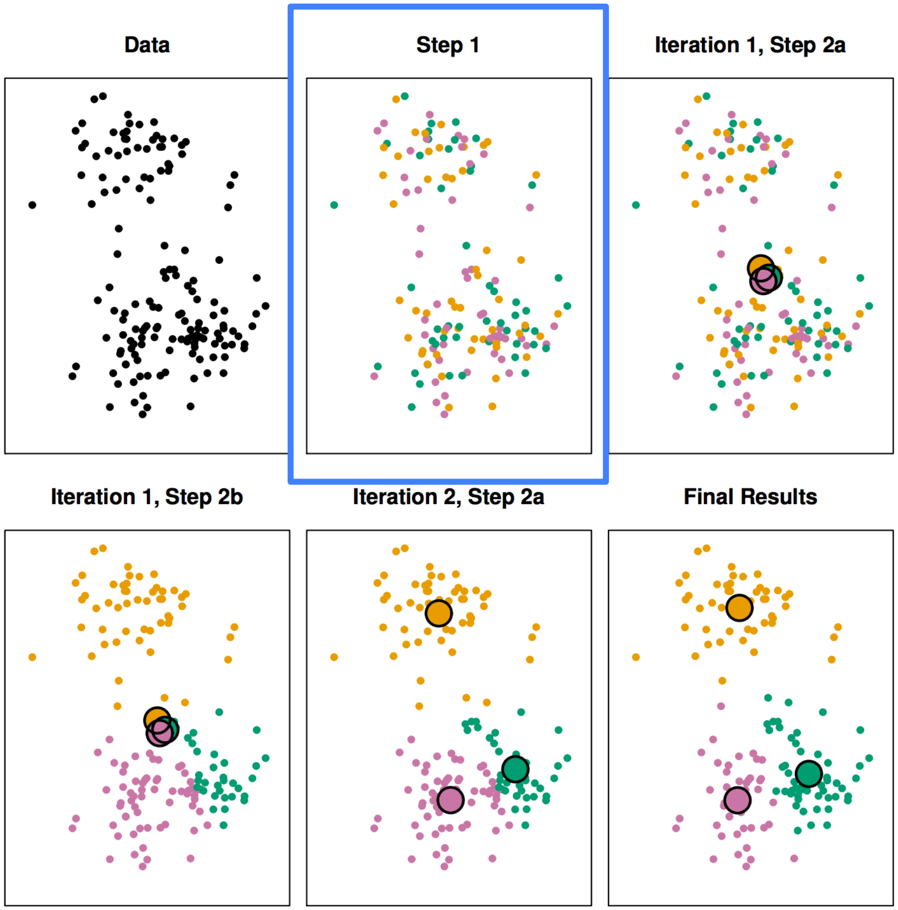
Idea: Want "within-cluster variation" to be small

Suppose: A fixed K, say K=3. Want to assign each of *n* data point to one of 3 clusters, such that "within-cluster variation" is smallest

- There are K^n possible choices! Pretty unwieldy



K-means algorithm



Finds local optimum!

Results depend on random initialization

Solution

Try multiple initializations and pick one with lowest

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

* Also could consider smarter initializations such as kmeans++ http://en.wikipedia. org/wiki/K-means%2B%2B

Choosing K

- No easy answer
- A fuzzy endeavor
  - May just want K similar groups
  - But more often, want something useful or interpretable that exposes some interesting aspect of data
    - Presence/absence of natural distinct groups
    - Descriptive statistics about groups
  - Ex. Are there certain segments of my market that tend to be alike?
    - Ex. middle-aged living in suburbs who log-in infrequently

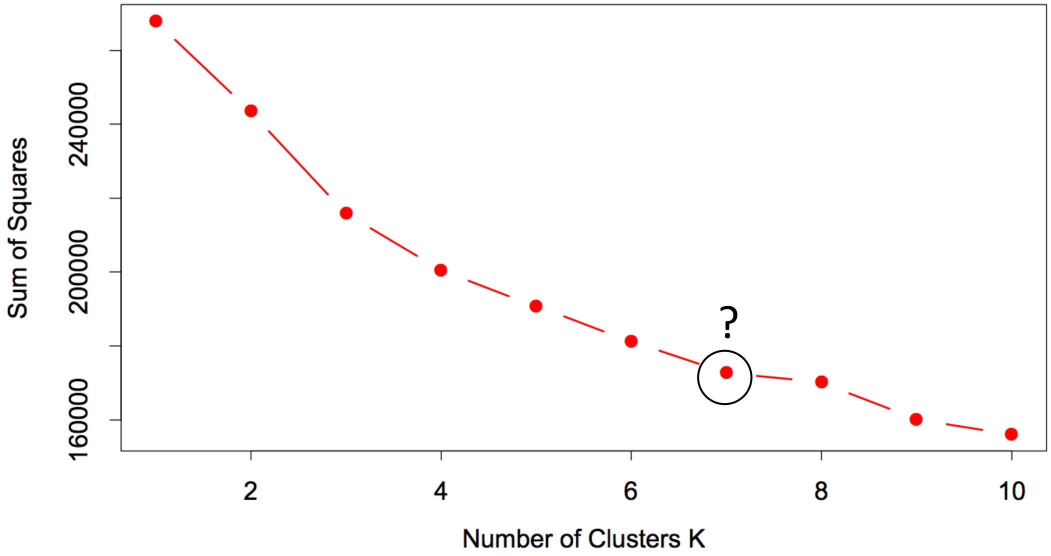
Choosing K – "Elbow" method

- Same Idea: Choose a number of clusters so that adding another cluster doesn't give us that much more

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}||x_i - x_{i'}||^2$$

Within Cluster Point Scatter

A natural loss function is the sum pairwise distances of the points within each cluster, summed over all clusters.


Sum of Squares
?
Number of Clusters K