Clustering images

Set of Images

©Carlos Guestrin 2005-2013
Clustering web search results

Some Data
K-means

1. Ask user how many clusters they'd like.
   \((\text{e.g. } k=5)\)

2. Randomly guess \(k\) cluster Center locations
**K-means**

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center “owns” a set of datapoints)
4. Each Center finds the centroid of the points it owns
K-means

1. Ask user how many clusters they’d like. (e.g. k=5)
2. Randomly guess k cluster center locations
3. Each datapoint finds out which center it’s closest to.
4. Each center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!

![K-means Diagram]

**K-means**

- Randomly initialize $k$ centers
  - $\mu^{(0)} = \mu_1^{(0)}, \ldots, \mu_k^{(0)}$
- **Classify**: Assign each point $j \in \{1, \ldots, m\}$ to nearest center:
  - $C(t)(j) \leftarrow \arg \min_i ||\mu_i - x_j||^2$
- **Recenter**: $\mu_i$ becomes centroid of its point:
  - $\mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j: C(j) = i} ||\mu - x_j||^2$
  - $\mu_i = \frac{\sum_{j: C(j) = i} x_j}{\text{num of points assigned to cluster } i}$

©Carlos Guestrin 2005-2013
What is K-means optimizing?

- Potential function $F(\mu, C)$ of centers $\mu$ and point allocations $C$:
  \[
  F(\mu, C) = \sum_{j=1}^{N} \| \mu_{C(j)} - x_j \|^2
  \]

- Optimal K-means:
  \[
  \min_{\mu} \min_{C} F(\mu, C)
  \]

Does K-means converge??? Part 1

- Optimize potential function:
  \[
  \min_{\mu} \min_{C} F(\mu, C) = \min_{\mu} \min_{C} \sum_{i=1}^{k} \sum_{j: C(j) = i} \| \mu_i - x_j \|^2
  \]

- Fix $\mu$, optimize $C$
  \[
  \min_{C \in \text{partition}} \sum_{j=1}^{N} \| \mu_{C(j)} - x_j \|^2 = \min_{C(1), \ldots, C(k)} \sum_{j=1}^{N} \| \mu_{C(j)} - x_j \|^2
  \]

©Carlos Guestrin 2005-2013
Does K-means converge??? Part 2

- Optimize potential function:
  \[ \min_{\mu} \min_{C} F(\mu, C) = \min_{\mu} \min_{C} \sum_{i=1}^{k} \sum_{j: C(j) = i} \| \mu_i - x_j \|^2 \]

- Fix \( C \), optimize \( \mu \)
  \[ \min_{\mu: \mu \in \mathbb{R}^d} \sum_{i=1}^{k} \sum_{j: C(j) = i} \| \mu_i - x_j \|^2 \]

  \[ \mu_i = \text{center of points in cluster } i \]
  \[ \mu_i = \text{average of points in cluster } i \]

Coordinate descent algorithms

- Want: \( \min_a \min_b F(a,b) \)
- Coordinate descent:
  - fix \( a \), minimize \( b \)
  - fix \( b \), minimize \( a \)
  - repeat
- Converges!!!
  - if \( F \) is bounded
  - to a (often good) local optimum
    - as we saw in applet (play with it!)
    - (For LASSO it converged to the optimum)
  - Random restarts help

- K-means is a coordinate descent algorithm!
(One) bad case for k-means

- Clusters may overlap
- Some clusters may be “wider” than others
Gaussians in $m$ Dimensions

$$P(x) = \frac{1}{(2\pi)^{m/2} ||\Sigma||^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]$$

Suppose You Have a Gaussian For Each Class

$$P(x \mid y = i) \propto \frac{1}{(2\pi)^{m/2} ||\Sigma||^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right]$$
Gaussian Bayes Classifier

- You have a Gaussian over $x$ for each class $y=i$:
  $$P(x \mid y = i) = \frac{1}{(2\pi)^{m/2} |\Sigma|^1/2} \exp\left(-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right)$$

But you need probability of class $y=i$ given $x$:

- Thank you Bayes Rule!!

$$P(y = i \mid x) = \frac{P(x \mid y = i)P(y = i)}{p(x)}$$

- Predicting wealth from age

- Prior: what fraction of points we expect to fall in cluster $i$ before we see $x$.
- Likelihood: $p(x \mid y = i)$.
Predicting wealth from age

Learning model year, mpg --> maker

\[ \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix} \]
General: $O(m^2)$

parameters

\begin{equation}
\Sigma = \begin{pmatrix}
\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\
\sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 
\end{pmatrix}
\end{equation}