K-Means

David Rosenberg, Brett Bernstein

New York University

April 25, 2017

Intro Question

Intro Question

Consider the following probability model for generating data.

1. Roll a weighted k-sided die to choose a label $z \in \{1, \ldots, k\}$. Let π denote the PMF for the die.
2. Draw $x \in \mathbf{R}^d$ randomly from the multivariate normal distribution $\mathcal{N}(\mu_z, \Sigma_z)$.

Solve the following questions.

1. What is the joint distribution of x, z given π and the μ_z, Σ_z values?
2. Suppose you were given the dataset $\mathcal{D} = \{(x_1, z_1), \ldots, (x_n, z_n)\}$. How would you estimate the die weightings, and the μ_z, Σ_z values?
3. How would you determine the label for a new datapoint x?

Intro Solution

1. The joint PDF/PMF is given by

$$p(x,z) = \pi(z) f(x; \mu_z, \Sigma_z)$$

where

$$f(x; \mu_z, \Sigma_z) = \frac{1}{\sqrt{|2\pi\Sigma_z|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

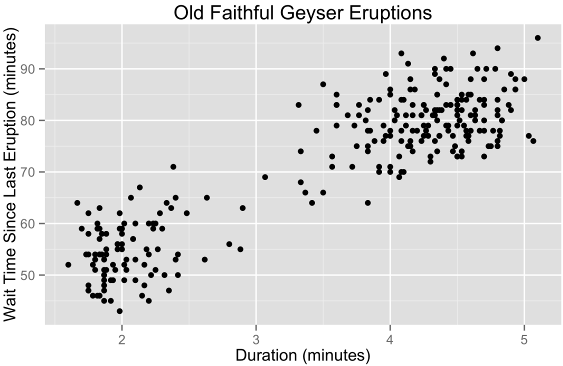
2. We could use maximum likelihood estimation. Our estimates are

$$\begin{aligned} n_z &= \textstyle\sum_{i=1}^n \mathbf{1}(z_i = z) \\ \hat{\pi}(z) &= \tfrac{n_z}{n} \\ \hat{\mu}_z &= \tfrac{1}{n_z} \textstyle\sum_{i:z_i=z} x_i \\ \hat{\Sigma}_z &= \tfrac{1}{n_z} \textstyle\sum_{i:z_i=z} (x_i - \hat{\mu}_z)(x_i - \hat{\mu}_z)^T. \end{aligned}$$

3. $\arg\max_z p(x,z)$

K-Means Clustering

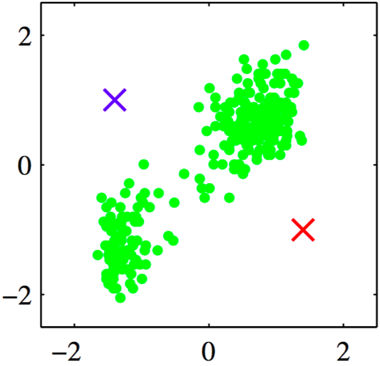
Example: Old Faithful Geyser



- Looks like two clusters.
- How to find these clusters algorithmically?

k-Means: By Example

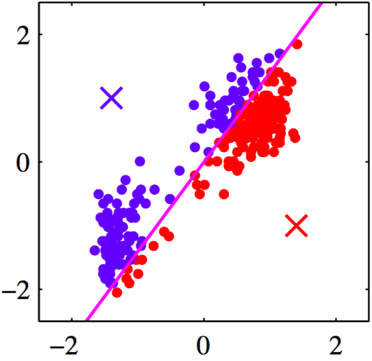
- Standardize the data.
- Choose two cluster centers.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(a).

k-means: by example

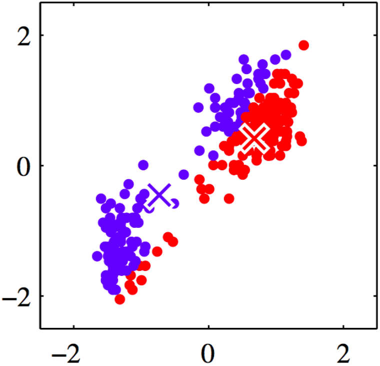
- Assign each point to closest center.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(b).

k-means: by example

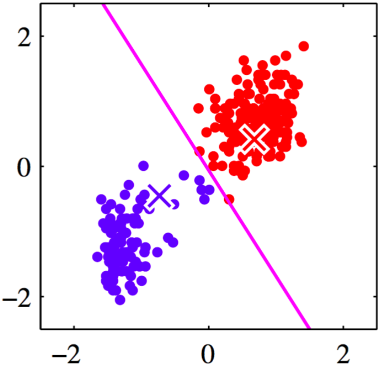
- Compute new class centers.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(c).

k-means: by example

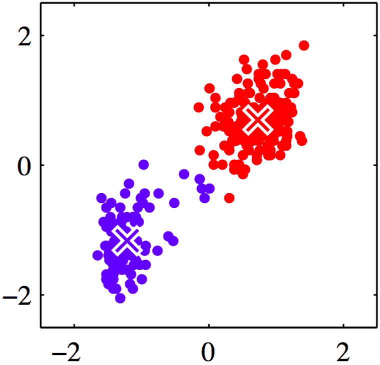
- Assign points to closest center.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(d).

k-means: by example

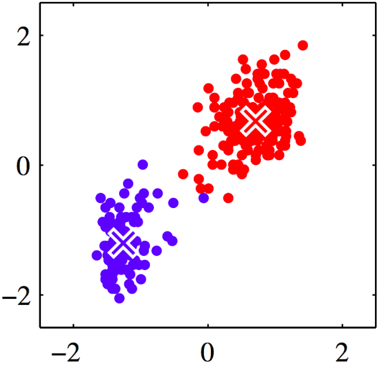
- Compute cluster centers.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(e).

k-means: by example

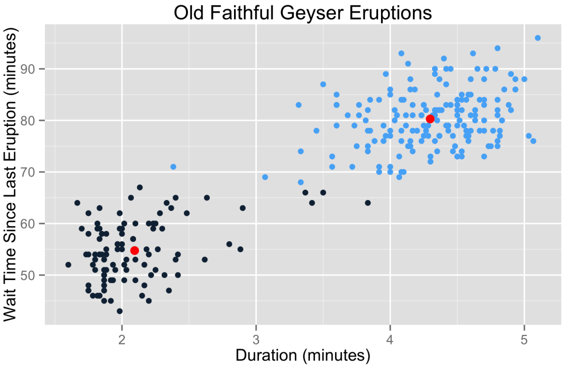
- Iterate until convergence.



From Bishop's *Pattern recognition and machine learning*, Figure 9.1(i).

k-Means Algorithm: Standardizing the data

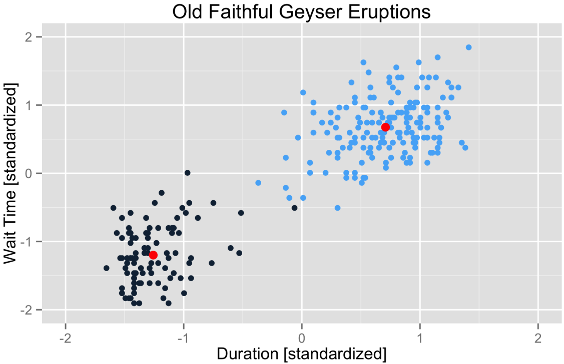
- Without standardizing:



- Blue and black show results of k-means clustering
- Wait time dominates the distance metric

k-Means Algorithm: Standardizing the data

- With standardizing:



- Note several points have been reassigned from black to blue cluster.

k-Means: Objective

- Let $x_1, \ldots, x_n$ denote the data points and $\mu_1, \ldots, \mu_k$ the cluster points.
- Define the objective ϕ by

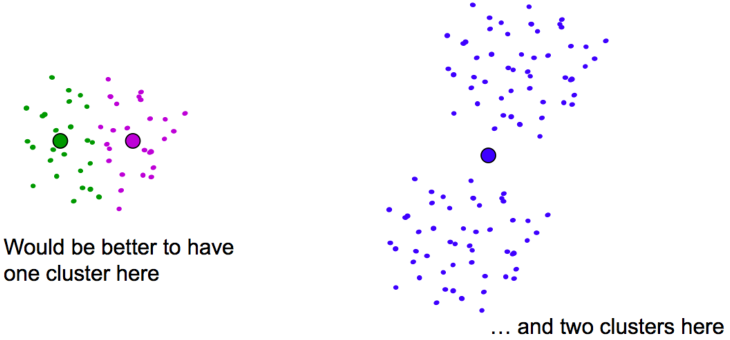
$$\phi(x, \mu) = \sum_{i=1}^{n} \|x_i - \mu_{c(x_i)}\|_2^2,$$

  where $\mu_{c(x_i)}$ is the cluster point associated to x_i.
- Then ϕ decreases at every round of k-means. Why?
- Selecting mean of all associated data points improves objective.
- Selecting closest cluster point for each data points improves objective.

k-Means: Failure Cases

k-Means: Suboptimal Local Minimum

- The clustering for $k = 3$ below is a local minimum, but suboptimal:



From Sontag's DS-GA 1003, 2014, Lecture 8.

k-Means++

- Improvement on k-means by controlling the random initialization of the cluster centers.
- Randomly choose first center amongst the data points.
- For each of the remaining $k-1$ centers:
    - Compute the distance from each data point to the closest already chosen center.
    - Randomly choose a point as the new center with probability proportional to its computed distance squared.
- If we let ϕ denote the total sum of squares distances from each point to the closest cluster, then k-means++ has

$$E[\phi] \leqslant 8(\log k + 2)\phi_{\text{OPT}},$$

where ϕ_{OPT} is from the optimal k-cluster assignment.