Statistical Machine Learning from Data

Classical Models

Samy Bengio

IDIAP Research Institute, Martigny, Switzerland, and
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
bengio@idiap.ch
http://www.idiap.ch/~bengio

February 8, 2006

Parametric or Not?

- The space $\mathcal{F}$ is often characterized to be parametric or not.
- Parametric: the space is very small, and characterized by a small number of parameters.
    - examples: a Gaussian distribution or a linear function
    - big prior on the solution
- Non-Parametric: the space is infinite, constrained only by the training data
    - examples: K nearest neighbors, Parzen Windows
    - small prior on the solution
- Semi-Parametric:
    - examples: most machine learning algorithms!
    - small prior on the solution, characterized by a large number of parameters

1. Histograms, K Nearest Neighbors and Parzen Windows
2. Maximum Likelihood and Bayes Decision
3. K-Means
4. Linear Regression

Histograms - Example

- For classification or regression: $z = (x, y)$
- Let x be a $k-$dimensional vector
- For each dimension d, divide the possible values x_d into m_d bins
- Example, for 14 training examples of input dimension 2:

	$x_1 < 5$	$5 \leq x_1 < 7$	$7 \leq x_1$
$x_2 =$ red	$y(0) = -3$ $y(1) = -4$ $y(2) = -2.8$	$y(3) = 2$ $y(4) = 1$	$y(5) = 3$ $y(6) = 4$ $y(7) = 2.8$ $y(8) = 2.5$
$x_2 =$ blue	$y(9) = -4.5$ $y(10) = -4$	$y(11) = 0.1$	$y(12) = 0.1$ $y(13) = 0.65$

Histograms - Training

- Model: compute average value (on the training set) of $\hat{y}$ corresponding to each bin:

	$x_1 < 5$	$5 \leq x_1 < 7$	$7 \leq x_1$
$x_2 =$ red	$\hat{y} = -3.3$	$\hat{y} = 1.5$	$\hat{y} = 3.1$
$x_2 =$ blue	$\hat{y} = 4.25$	$\hat{y} = 0.1$	$\hat{y} = 0.38$

- Test: given a new example x, select the corresponding bin and output the associated $\hat{y}$
- Can be extended to classification.
- Capacity controlled by the total number of bins.
- Total number of bins $= \displaystyle\prod_{d=1}^{k} m_d$

Problem: The Curse of Dimensionality (1)

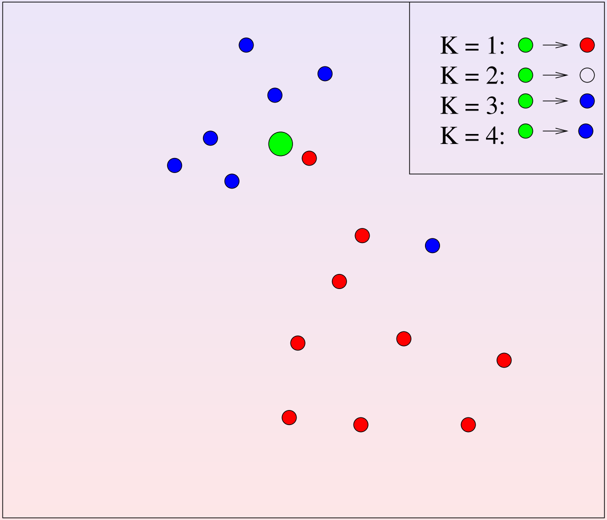
Combinatorial Explosion

- What happens when the number of input dimensions grows?
- The number of bins grows exponentially faster!
- Most bins will get no representative training example
- How can we estimate a new example that is in one of those bins????
- In fact, even the bins with some training examples are probably not correctly estimated...

K Nearest Neighbors

- Very simple method, no training necessary
- Needed:
  - a training set $D_n = \{z_1, z_2, \cdots, z_n\}$ with $z_i = (x_i, y_i)$
  - a distance function $L(x_1, x_2)$. For instance, $(x_1 - x_2)^2$
  - a parameter K
- For each test point x
  - select in D_n the K examples that are nearest to x according to $L(x, x_i)$ and keep their index (from D_n) in $\{s_1, \cdots, s_K\}$
  - decision:
    - regression: $\hat{y} = \frac{1}{K} \sum_{i=1}^{K} y_{s_i}$
    - classification: $\hat{y} = \text{sign}\left(\frac{1}{K} \sum_{i=1}^{K} y_{s_i}\right)$
- Capacity controlled by K.

K-NN (Graphical View)



KNN - Some Remarks

- What does it mean to be nearest to an example?
- Often used metric: Euclidean distance, or l^2-norm

$$d = \sqrt{\sum_i (x_i - t_i)^2}$$

- For KNN, $\sqrt{\cdot}$ is not necessary
- How to select K ???
- Reminder: K controls the capacity...
- Hence, we can use a model selection technique

Distances and the Curse of Dimensionality

- Consider a regular grid of b bins per dimension in a d-dim hypercube.

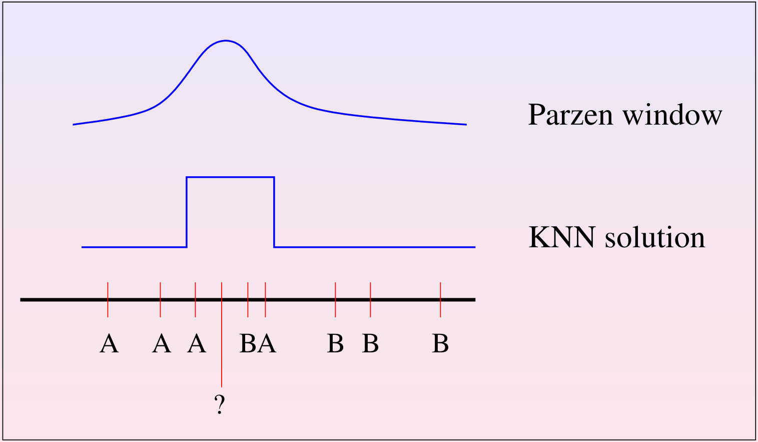
  We have b^d bins.

- How many bins are not on the surface of the hypercube? consider uniform data.

$$\left(\frac{b-2}{b}\right)^d \text{ chances of being in the center.} \quad \longrightarrow 0$$

- When d is high, all data lie on the surface!
- How many points of the training set can be on the same surface? As d grows, less than one on average!
- Each point is thus far from all the others...
- Hence, all methods based on Euclidean distance are bound to work on small dimensions only.

KNN versus Parzen Windows



Parzen Windows

- Very simple method, no training necessary
- Needed:
  - a training set $D_n = \{z_1, z_2, \cdots, z_n\}$ with $z_i = (x_i, y_i)$
  - a kernel function $K(x_1, x_2)$. For instance, $\exp(-\frac{||x_1 - x_2||^2}{2\sigma^2})$
- For each test point x (or z for density estimate)
  - decision:
    - regression: $\hat{y}_r = \dfrac{\sum_{i=1}^{n} y_i K(x, x_i)}{\sum_{i=1}^{n} K(x, x_i)}$
    - classification: $\hat{y} = \text{sign}(\hat{y}_r)$
    - density estimate: $\hat{p}(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} K(x, x_i)$
- Capacity controlled by σ

Maximum Likelihood for Density Estimation

- Given a set of examples $D_n = \{z_1, z_2, \cdots, z_n\}$
- Objective: find a distribution $p(z)$ that maximizes the likelihood of future data
- Select a set of distributions $p(z|\theta)$ with parameters θ.
- The likelihood of D_n (all examples are iid):

$$\mathcal{L}(D_n|\theta) = \prod_{i=1}^{n} p(z_i|\theta)$$

Hence we search for:

$$\theta^* = \arg\max_\theta \prod_{i=1}^{n} p(z_i|\theta) = \arg\max_\theta \sum_{i=1}^{n} \log p(z_i|\theta)$$

going into the log domain.

Maximum Likelihood for Gaussians

- Family of one-dimensional Gaussians with $\theta = \{\mu, \sigma\}$

$$\hat{p}(z|\theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

- The log likelihood for a set of n data is thus:

$$l = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right)$$

- In order to find the maximum likelihood solution, we need to set

$$\frac{\partial l}{\partial \theta} = 0$$

for parameters $\theta = \{\mu, \sigma\}$

Maximum Likelihood Solution - Means

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \mu} &= \sum_{i=1}^{n}\frac{z_i-\mu}{\sigma^2} = \sum_{i=1}^{n}\frac{z_i}{\sigma^2} - \sum_{i=1}^{n}\frac{\mu}{\sigma^2} \\
0 &= \sum_{i=1}^{n}\frac{z_i}{\sigma^2} - \frac{n\mu}{\sigma^2}
\end{aligned}
$$

$$
\Longrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} z_i
$$

Maximum Likelihood Solution - Means

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \mu} &= \sum_{i=1}^{n} \frac{z_i-\mu}{\sigma^2} = \sum_{i=1}^{n}\frac{z_i}{\sigma^2} - \sum_{i=1}^{n}\frac{\mu}{\sigma^2} \\
0 &= \sum_{i=1}^{n}\frac{z_i}{\sigma^2} - \frac{n\mu}{\sigma^2}
\end{aligned}
$$

$$
\Longrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} z_i
$$

Maximum Likelihood Solution - Means

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \mu} &= \sum_{i=1}^{n} \frac{z_i-\mu}{\sigma^2} = \sum_{i=1}^{n} \frac{z_i}{\sigma^2} - \sum_{i=1}^{n} \frac{\mu}{\sigma^2} \\
0 &= \sum_{i=1}^{n} \frac{z_i}{\sigma^2} - \frac{n\mu}{\sigma^2}
\end{aligned}
$$

$$
\Longrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} z_i
$$

Maximum Likelihood Solution - Means

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n} \frac{(z_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \mu} &= \sum_{i=1}^{n} \frac{z_i-\mu}{\sigma^2} = \sum_{i=1}^{n} \frac{z_i}{\sigma^2} - \sum_{i=1}^{n} \frac{\mu}{\sigma^2} \\
0 &= \sum_{i=1}^{n} \frac{z_i}{\sigma^2} - \frac{n\mu}{\sigma^2}
\end{aligned}
$$

$$
\Longrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} z_i
$$

Maximum Likelihood Solution - Means

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l}{\partial \mu} &= \sum_{i=1}^{n}\frac{z_i-\mu}{\sigma^2} = \sum_{i=1}^{n}\frac{z_i}{\sigma^2} - \sum_{i=1}^{n}\frac{\mu}{\sigma^2} \\
0 &= \sum_{i=1}^{n}\frac{z_i}{\sigma^2} - \frac{n\mu}{\sigma^2}
\end{aligned}
$$

$$
\Longrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} z_i
$$

Maximum Likelihood Solution - Variances

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^4} = 0
$$

$$
\begin{aligned}
\Longrightarrow \quad \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^4} &= \frac{n}{2\sigma^2} \\
\frac{1}{n}\sum_{i=1}^{n}(z_i-\mu)^2 &= \sigma^2
\end{aligned}
$$

Maximum Likelihood Solution - Variances

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^4} = 0
$$

$$
\begin{aligned}
\Longrightarrow \quad \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^4} &= \frac{n}{2\sigma^2} \\
\frac{1}{n}\sum_{i=1}^{n}(z_i-\mu)^2 &= \sigma^2
\end{aligned}
$$

Maximum Likelihood Solution - Variances

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^4} = 0
$$

$$
\begin{aligned}
\Longrightarrow \quad \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^4} &= \frac{n}{2\sigma^2} \\
\frac{1}{n}\sum_{i=1}^{n}(z_i-\mu)^2 &= \sigma^2
\end{aligned}
$$

Maximum Likelihood Solution - Variances

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^4} = 0
$$

$$
\Longrightarrow \quad \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^4} = \frac{n}{2\sigma^2}
$$

$$
\frac{1}{n}\sum_{i=1}^{n}(z_i-\mu)^2 = \sigma^2
$$

Maximum Likelihood Solution - Variances

$$
\begin{aligned}
l &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^2}
\end{aligned}
$$

$$
\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^4} = 0
$$

$$
\begin{aligned}
\Longrightarrow \quad \sum_{i=1}^{n}\frac{(z_i-\mu)^2}{2\sigma^4} &= \frac{n}{2\sigma^2} \\
\frac{1}{n}\sum_{i=1}^{n}(z_i-\mu)^2 &= \sigma^2
\end{aligned}
$$

Bayes Decision

- Classification: $z = (x, y) \in \mathbb{R}^d \times \{-1, 1\}$
- Given: true posterior distribution $P(Y = y|X = x)$
- It can be shown that the decision

$$\hat{y} = \arg \max_{i \in \{1,-1\}} P(Y = i|X = x)$$

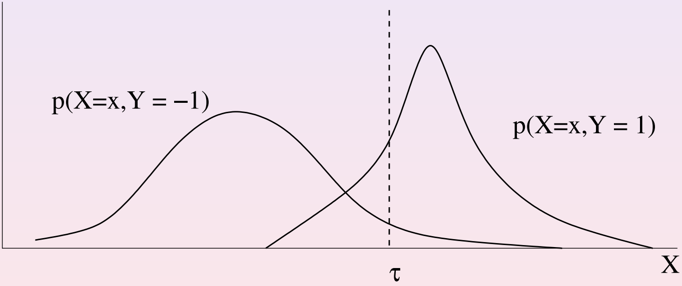
  is optimal in the sense that it minimizes the number of classification errors.
- This decision corresponds to the class maximum a posteriori (MAP) criterion

Why Class MAP Minimizes Error?

$$
\begin{aligned}
\hat{y} &= \arg \max_{i \in \{1,-1\}} P(Y = i | X = x) \\
&= \arg \max_{i \in \{1,-1\}} \frac{p(X = x | Y = i) \cdot P(Y = i)}{p(X = x)} \\
&= \arg \max_{i \in \{1,-1\}} p(X = x | Y = i) \cdot P(Y = i) \\
&= \arg \max_{i \in \{1,-1\}} p(X = x, Y = i)
\end{aligned}
$$

Why Class MAP Minimizes Error?

- Let us select a threshold for all our decisions $X = \tau$.



Why Class MAP Minimizes Error?

- The ratio of errors we make can be decomposed into two terms:
  - when $X > \tau$ but $Y = -1$

$$= p(X > \tau, Y = -1) = \int_{x>\tau} p(X = x, Y = -1)dx$$

  - when $X < \tau$ but $Y = 1$

$$= p(X < \tau, Y = 1) = \int_{x<\tau} p(X = x, Y = 1)dx$$

- Which τ corresponds to the minimum error?

$$\tau^\star = \arg\min_{\tau} p(X > \tau, Y = -1) + p(X < \tau, Y = 1)$$

which happens exactly when

$$p(X = \tau, Y = -1) = p(X = \tau, Y = 1)$$

Bayes Classifiers

- Goal: take the decision based on the MAP criterion:

$$\hat{y} = \arg \max_{i \in \{1,-1\}} p(X = x | Y = i) \cdot P(Y = i)$$

- Hence, you need to estimate:
  - the conditional density $p(X = x|Y = i)$ for each class i
  - the class prior $P(Y = i)$ for each class i
- Good: each class is estimated independently
- Bad: you learn unnecessary relations
- This technique is nevertheless often used in speech processing

Naive Bayes Classifiers

- Classification decision according to a Bayes Classifier:

$$\hat{y} = \arg \max_{i \in \{1,-1\}} p(X = x | Y = i) \cdot P(Y = i)$$

- $P(Y = i)$ can be estimated by counting in the training set.
- We need a way to represent $p(X = x | Y = i)$.
- Let us suppose that $X \in \mathbb{R}^d$ AND all X_j are independent...
- Hence, the Naive Bayes model assumes:

$$p(X = x | Y = i) = p(X_1 = x_1, \ldots, X_d = x_d | Y = i) = \prod_{j=1}^{d} p(X_j = x_j | Y = i)$$

- So, the Naive Bayes Classifier becomes:

$$\hat{y} = \arg \max_{i \in \{1,-1\}} P(Y = i) \cdot \prod_{j=1}^{d} p(X_j = x_j | Y = i)$$

Clustering by K-Means

- Given a set of examples $D_n = \{z_1, z_2, \cdots, z_n\}$
- Search for K prototypes μ_k of disjoint subsets S_k of D_n in order to minimize

$$L = \sum_{k=1}^{K} \sum_{j \in S_k} \|z_j - \mu_k\|^2$$

where μ_k is the mean of the examples in subset S_k:

$$\mu_k = \frac{1}{|S_k|} \sum_{j \in S_k} z_j$$

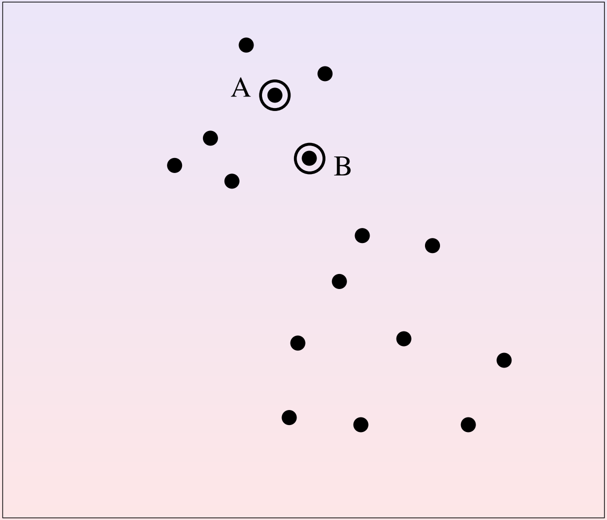
- We could use any distance, not just the Euclidean distance...

Batch and Stochastic K-Means

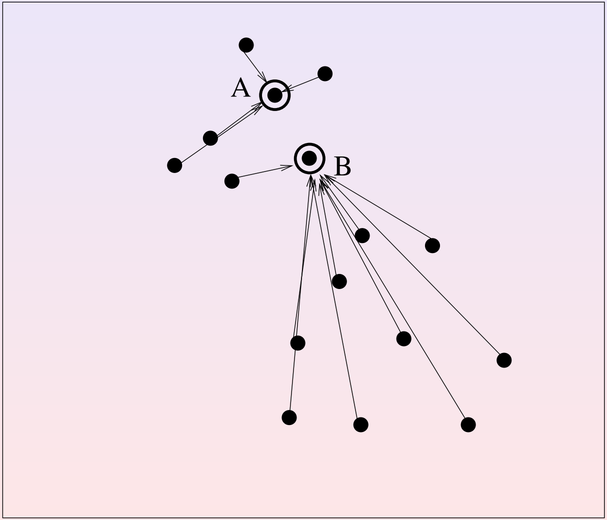
- Initialization: select randomly K examples z_j in D_n as initial values of each μ_k
- At each batch iteration:
  - For each prototype μ_k, put in the emptied set S_k the examples of D_n that are closer to μ_k than to any other $\mu_{j \neq k}$.
  - Re-compute the value of each μ_k as the average of the examples in S_k.
- The algorithm stops when no prototype moves anymore.
- It can be shown that the K-Means criterion will never increase.
- A stochastic version of K-Means can also be derived: given a small η, for each example z_j move the nearest μ_k as follows:

$$\mu_k = \mu_k + \eta(z_j - \mu_k)$$

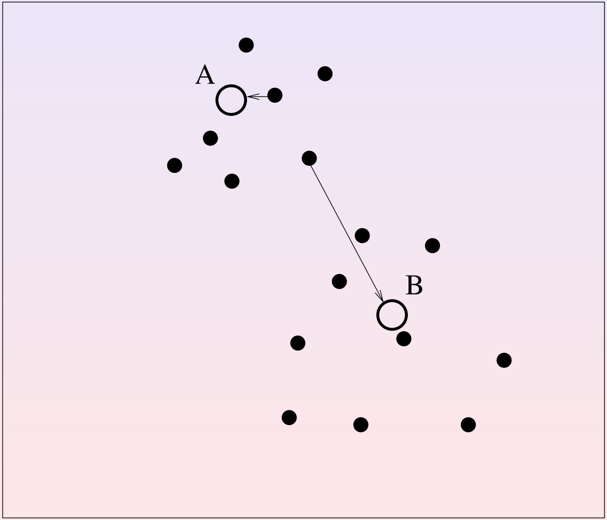
K-Means (Graphical View 1)



K-Means (Graphical View 2)



K-Means (Graphical View 3)



Convergence of K-Means

- Let μ^t be the set of clusters at time t
- Let $s(z_i, \mu^t) = \arg\min_k \|z_i - \mu_k^t\|^2$ the best cluster in μ^t for z_i.
- Let us rewrite $L(\mu^t) = \sum_{i=1}^{n} \|z_i - \mu^t_{s(z_i,\mu^t)}\|^2$
- We want to show that

$$L(\mu^{t+1}) - L(\mu^t) \leq 0$$

- Let $\mu_k^{t+1} = \frac{1}{|S_k|} \sum_{i \in S_k} z_i$ with S_k the set of z_i assigned to μ_k
- Let $Q(\mu^{t+1}, \mu^t) = \sum_{i=1}^{n} \|z_i - \mu^{t+1}_{s(z_i,\mu^t)}\|^2$

Convergence of K-Means

- Let μ^t be the set of clusters at time t
- Let $s(z_i, \mu^t) = \arg\min_k \|z_i - \mu_k^t\|^2$ the best cluster in μ^t for z_i.
- Let us rewrite $L(\mu^t) = \sum_{i=1}^{n} \|z_i - \mu^t_{s(z_i,\mu^t)}\|^2$
- We want to show that

$$L(\mu^{t+1}) - L(\mu^t) \leq 0$$

- Let $\mu_k^{t+1} = \frac{1}{|S_k|} \sum_{i \in S_k} z_i$ with S_k the set of z_i assigned to μ_k
- Let $Q(\mu^{t+1}, \mu^t) = \sum_{i=1}^{n} \|z_i - \mu^{t+1}_{s(z_i,\mu^t)}\|^2$

Convergence of K-Means

- Let μ^t be the set of clusters at time t
- Let $s(z_i, \mu^t) = \arg\min_k \|z_i - \mu_k^t\|^2$ the best cluster in μ^t for z_i.
- Let us rewrite $L(\mu^t) = \sum_{i=1}^{n} \|z_i - \mu^t_{s(z_i,\mu^t)}\|^2$
- We want to show that

$$L(\mu^{t+1}) - L(\mu^t) \leq 0$$

- Let $\mu_k^{t+1} = \frac{1}{|S_k|} \sum_{i \in S_k} z_i$ with S_k the set of z_i assigned to μ_k
- Let $Q(\mu^{t+1}, \mu^t) = \sum_{i=1}^{n} \|z_i - \mu^{t+1}_{s(z_i,\mu^t)}\|^2$

Convergence of K-Means

$$L(\mu^{t+1}) - L(\mu^t) = L(\mu^{t+1}) - Q(\mu^{t+1}, \mu^t) + Q(\mu^{t+1}, \mu^t) - L(\mu^t)$$

$$L(\mu^{t+1}) - Q(\mu^{t+1}, \mu^t) = \sum_{i=1}^{n} \left(\|z_i - \mu^{t+1}_{s(z_i, \mu^{t+1})}\|^2 - \|z_i - \mu^{t+1}_{s(z_i, \mu^t)}\|^2 \right) \leq 0$$

$$Q(\mu^{t+1}, \mu^t) - L(\mu^t) = \sum_{i=1}^{n} \left(\|z_i - \mu^{t+1}_{s(z_i, \mu^t)}\|^2 - \|z_i - \mu^{t}_{s(z_i, \mu^t)}\|^2 \right) \leq 0$$

$\Longrightarrow$

$$L(\mu^{t+1}) - L(\mu^t) \leq 0$$

Convergence of K-Means

$$L(\mu^{t+1}) - L(\mu^t) = L(\mu^{t+1}) - Q(\mu^{t+1}, \mu^t) + Q(\mu^{t+1}, \mu^t) - L(\mu^t)$$

$$L(\mu^{t+1}) - Q(\mu^{t+1}, \mu^t) = \sum_{i=1}^{n} \left(\|z_i - \mu^{t+1}_{s(z_i,\mu^{t+1})}\|^2 - \|z_i - \mu^{t+1}_{s(z_i,\mu^t)}\|^2 \right) \leq 0$$

$$Q(\mu^{t+1}, \mu^t) - L(\mu^t) = \sum_{i=1}^{n} \left(\|z_i - \mu^{t+1}_{s(z_i,\mu^t)}\|^2 - \|z_i - \mu^{t}_{s(z_i,\mu^t)}\|^2 \right) \leq 0$$

$$\Longrightarrow$$

$$L(\mu^{t+1}) - L(\mu^t) \leq 0$$

Convergence of K-Means

$$L(\mu^{t+1}) - L(\mu^t) = L(\mu^{t+1}) - Q(\mu^{t+1}, \mu^t) + Q(\mu^{t+1}, \mu^t) - L(\mu^t)$$

$$L(\mu^{t+1}) - Q(\mu^{t+1}, \mu^t) = \sum_{i=1}^{n} \left(\|z_i - \mu^{t+1}_{s(z_i, \mu^{t+1})}\|^2 - \|z_i - \mu^{t+1}_{s(z_i, \mu^t)}\|^2 \right) \leq 0$$

$$Q(\mu^{t+1}, \mu^t) - L(\mu^t) = \sum_{i=1}^{n} \left(\|z_i - \mu^{t+1}_{s(z_i, \mu^t)}\|^2 - \|z_i - \mu^{t}_{s(z_i, \mu^t)}\|^2 \right) \leq 0$$

$\Longrightarrow$

$$L(\mu^{t+1}) - L(\mu^t) \leq 0$$

Convergence of K-Means

$$L(\mu^{t+1}) - L(\mu^t) = L(\mu^{t+1}) - Q(\mu^{t+1}, \mu^t) + Q(\mu^{t+1}, \mu^t) - L(\mu^t)$$

$$L(\mu^{t+1}) - Q(\mu^{t+1}, \mu^t) = \sum_{i=1}^{n} \left(\|z_i - \mu^{t+1}_{s(z_i, \mu^{t+1})}\|^2 - \|z_i - \mu^{t+1}_{s(z_i, \mu^t)}\|^2 \right) \leq 0$$

$$Q(\mu^{t+1}, \mu^t) - L(\mu^t) = \sum_{i=1}^{n} \left(\|z_i - \mu^{t+1}_{s(z_i, \mu^t)}\|^2 - \|z_i - \mu^{t}_{s(z_i, \mu^t)}\|^2 \right) \leq 0$$

$\Longrightarrow$

$$L(\mu^{t+1}) - L(\mu^t) \leq 0$$

K-Means - Some Remarks

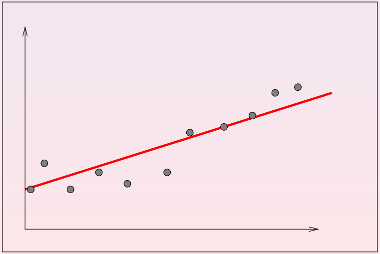
- As for KNN, we can change the metric
- For instance, we can normalize the data
- How to select K ???
- Reminder: as for KNN, K controls the capacity...
- Hence, we can use a model selection technique
- Note: K-Means is quite sensitive to initialization. Other heuristics exist, or you can retrain many times...
- Application: feature extraction

  represent each example z by the index of the closest prototype

1. Histograms, K Nearest Neighbors and Parzen Windows
2. Maximum Likelihood and Bayes Decision
3. K-Means
4. Linear Regression

Linear Regression

- We have a set of training examples $D_n = \{z_1, z_2, \cdots, z_n\}$
- With $z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$
- Linear function space: $\hat{y} = w \cdot x + b$ with parameters (w, b)
- Loss function: $L(y, \hat{y}) = (y - \hat{y})^2$



Solving Linear Regression

- The total error is as follows:

$$
\begin{aligned}
C &= \sum_i L(y, \hat{y}) \\
&= \sum_i (y_i - \hat{y}_i)^2 \\
&= \sum_i (y - w \cdot x_i - b)^2
\end{aligned}
$$

- We need to set simultaneously $\frac{\partial C}{\partial w}$ and $\frac{\partial C}{\partial b}$ to 0.
- For easier mathematical derivation $\longrightarrow$ matrix notation

Solving Linear Regression by Matrix Inversion

- Let $r_i = [x_i \;\; 1]$ the input vector of example i augmented by the value 1.
- Let R be the $(n \times (d+1))$ matrix of vectors r_i.
- Let Y be the $(n \times 1)$ target matrix.
- Let $v = [w \;\; b]$ be the $(d+1)$-dim vector concatenating w and b.
- The total cost is:

$$\begin{aligned} C &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y - (w \cdot x_i + b \cdot 1))^2 \\ &= (Y - Rv)'(Y - Rv) \end{aligned}$$

Solution of the Linear Regression Problem

- The cost:

$$C = (Y - Rv)'(Y - Rv)$$

- Its minimum should satisfy:

$$\frac{\partial C}{\partial v} = 0$$

- Let us solve:

$$\begin{aligned} \frac{\partial C}{\partial v} &= -2R'(Y - Rv) = 0 \\ \text{Hence: } \hat{v} &= (R'R)^{-1}R'Y \end{aligned}$$