Lec 10

Tuesday, October 1, 2019 10:51

Recap: PCA

Auto-encoder view of PCA

encoder $e : \mathbb{R}^p \to \mathbb{R}^q$

$z = e(x) = A^T x$

decoder $d : \mathbb{R}^q \to \mathbb{R}^p$

$\hat{x} = d(z) = Az$

Where $A \in \mathbb{R}^{p \times q}$, $A^T A = I$

Soln : PCA : get $A = V_q$

the first q cols of V in the SVD $X = U \Sigma V^T$

$\Updownarrow$

for minimizing

$\sum_i \|\hat{x}_i - x_i\|_2^2$

$\Updownarrow$

$\sum_i \|d(e(x_i)) - x_i\|_2^2$

PCA for Non-Centered data

decoder: $\hat{x} = d(z) = Az + a$

encoder: $z = e(x) = A^T(x - a)$

Best A, a are:

$a = \bar{x} = \frac{1}{n} \sum_i x_i$

$A = V_q$ = first q cols from the SVD of $X - \bar{x} = U \Sigma V^T$

where $(X - \bar{x})_i \overset{i\text{th row}}{=} x_i - \bar{x}$

= first q eigenvectors from the eigendecomp of

$\frac{1}{n}(X - \bar{x})^T (X - \bar{x})$

Sample Covariance matrix of the data

Clustering

Assign data points to finitely many, $k \in \mathbb{N}$, clusters

One perspective: find clusters in the data

Auto-encoder perspective:

dimensionality reduction w/

$z = e(x) \in \{1, \dots, k\}$

$\hat{x} = d(z) = \mu_z$

with dictionary $\{\mu_1, \dots, \mu_k\}$

K-means

Tries to assign datapts to clusters s.t. the within-cluster distances are small

Data: $\underline{X} \in \mathbb{R}^{n \times p}$

clusters: $k \in \mathbb{N}$

Let $C(i) = 1, \dots, k$ $\quad C : \{1, \dots, n\} \to \{1, \dots, k\}$

indicate the assignment of pt i to a cluster

Quality of a clustering C is defined as the within-cluster diffs:

$$W(C) = \sum_{i=1}^n \|x_i - \mu_{C(i)}\|_2^2$$

Where $\mu_j = \frac{1}{n_j} \sum_{i : C(i) = j} x_i$

$n_j = \sum_{i : C(i) = j} 1$

$$= \sum_{j=1}^k \sum_{i : C(i) = j} \|x_i - \mu_j\|_2^2$$

$$= \frac{1}{2} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{i: C(i)=j} \sum_{i': C(i')=j} \|x_i - x_{i'}\|_2^2$$

Want best C in terms of $W(C)$

How many C's are there?

\# ways to assign n things to k buckets

= Stirling # of 2nd kind

= HUGE!

e.g. for $k=4$, $S(10,4) = 34105$

$S(19,4) > 10^{10}$

K-means: greedy iterative approach to this hard optim problem

Start w/ some initial clustering C_0

For $t = 1, 2, \ldots$:

1. Compute the cluster means for C_{t-1}

$$\mu_j^{(t-1)} = \frac{1}{n_j^{(t-1)}} \sum_{i: C_{t-1}(i)=j} x_i \quad \forall j = 1, \ldots, k$$

2. Reassign each pt x_i to its closest center

$$C_t(i) = \underset{j=1,\ldots,k}{\operatorname{argmin}} \|x_i - \mu_j^{(t-1)}\|_2^2$$

3. Repeat

Obs 1: Different initializations (i.e. C_0)

lead to different solns.

Soln: Try diff (random) starts
& pick the result
w/ best W(C)

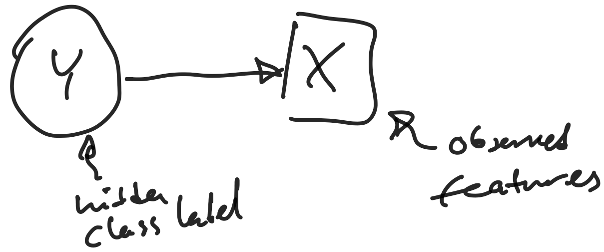
Obs 2: K-means will terminate/converge/stop moving
in a finite # of iterations.
Might not be the global opt (see obs 1)

Soft clustering

K-means alg assign each pt to
exactly one cluster
— hard clustering

Sometimes not clear that there's
a clear cut distinction into clusters

Soft clustering: assign % membership



Like we got a supervised learning data set
but Y col dropped (even in training)

Fit Gaussian Mixture Model
W/ EM algorithm

Suppose K=2
GMM: [illegible] distrib as follows

GMM: [illegible]

$$X_0 \sim N(\mu_0, \sigma_0^2) \qquad X_1 \sim N(\mu_1, \sigma_1^2)$$

$$Y \sim Ber(\pi)$$

$$X = (1-Y)X_0 + YX_1 = \begin{cases} X_0 & Y=0 \\ X_1 & Y=1 \end{cases}$$

Or: $(X \mid Y=y) \sim N(\mu_y, \sigma_y^2)$

Approach: find $\Theta = \{\pi, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2\}$

that best fit my data

(i.e., maximize the likelihood)

& then we'll use $\hat{\theta}$ to

compute $\hat{\mathbb{P}}_{\hat{\theta}}(Y=y \mid X)$ — our soft cluster membership in cluster y

$$\mathbb{P}(Y=1 \mid X=x) = \frac{\mathbb{P}(X=x \mid Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x \mid Y=1)\mathbb{P}(Y=1) + \mathbb{P}(X=x \mid Y=0)\mathbb{P}(Y=0)}$$

$$= \frac{\pi\, \varphi\left(\frac{x-\mu_1}{\sigma_1}\right)}{\pi\, \varphi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-\pi)\, \varphi\left(\frac{x-\mu_0}{\sigma_0}\right)}$$

= cluster responsibility

(soft membership to cluster 1)

Fitting this w/ EM

Given observation $X_1, \dots, X_n$ the log-lik is

$$\ell(\theta; \underline{X}) = \sum_{i=1}^n \log\left((1-\pi)\, \varphi\left(\frac{X_i-\mu_0}{\sigma_0}\right) + \pi\, \varphi\left(\frac{X_i-\mu_1}{\sigma_1}\right)\right)$$

We want θ to max $\ell(\theta; X)$ (or min $-\ell(\theta; X)$)

Actually: hard optim problem