Lec 9

Thursday, September 26, 2019 10:53

Recap: density estimation

See observations $Y_1, \dots, Y_n$

drawn from distrib w/ density f

Want to estimate f

KDE (kernel density estimator)

$$\hat{f}_n^{KDE}(y) = \frac{1}{n} \sum_{i=1}^n K_\lambda (y - Y_i)$$

$$K_\lambda(u) = \frac{1}{\lambda} \varphi(u/\lambda)$$

$$\varphi(u) = \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}\|u\|_2^2}$$

KDE → kernel regression

$$\mu(x) = \mathbb{E}[Y \mid X = x] = \int y \, f_{Y|X=x}(y) \, dy$$

$$= \int y \frac{f_{Y,X}(y,x)}{f_X(x)} dy$$

$$= \frac{\int y \, f_{Y,X}(y,x) \, dy}{f_X(x)}$$

$$= \frac{\int y f_{Y,X}(y,x) \, dy}{\int f_{Y,X}(y,x) \, dy}$$

All we have to do is estimate the density $f_{Y,X}(y,x)$

If we use the estimate $\hat{f}_{Y,X}^{KDE}(y,x)$

we get kernel regression

$$\hat{\mu}_n(x) = \frac{\sum_{i=1}^n K_\lambda(x_i - x)\, Y_i}{\sum_{i=1}^n K_\lambda(x_i - x)}$$

KDE → kernel classification

Let $Y = \begin{cases} +1 \\ 0 \end{cases}$

$$\mathbb{P}(Y=1 \mid X=x) \geq .5 \iff \underbrace{\mathbb{E}[Y \mid X=x]}_{\text{apply kernel regression}} \geq .5$$

Kernel classifier declare 1 iff

$$\frac{\sum_{i=1}^n K_\lambda(x_i - x) Y_i}{\sum_{i=1}^n K_\lambda(x_i - x)} \geq .5$$

$$= \frac{\sum_{i=1}^n K_\lambda(x_i - x) Y_i}{\sum_{i=1}^n K_\lambda(x_i - x) Y_i + \sum_{i=1}^n K_\lambda(x_i - x)(1 - Y_i)}$$

$\Downarrow$

declare 1 iff

$$\hat{\pi}_1 \hat{f}_1(x) \geq \hat{\pi}_0 \hat{f}_0(x)$$

$\hat{f}_1$: KDE of X among pos ex

$\hat{f}_0$: KDE of X among neg ex

Eigen- & singular value decompositions

For a square matrix $A \in \mathbb{R}^{p \times p}$

$\lambda \in \mathbb{C}$, $v \in \mathbb{C}^p$ are eigen-value/vector pair of A

if $Av = \lambda v$

If A is symmetric $(A = A^T)$
then all of its e-vals are real
Any square symmetric matrix can be diagonalized

$$A = \sum_{i=1}^{p} \lambda_i v_i v_i^T$$

s.t. $\lambda_i \in \mathbb{R}$

$$v_i^T v_j = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{o/w} \end{cases}$$ ← known as an orthonormal set of vectors

$$A = V \Lambda V^T \qquad \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

$$VV^T = I_{p \times p} = V^T V$$

known as eigendecomposition of A

- If all of A's e-vals are nonneg
  then A is called positive semidefinite (PSD)
  and we can define

$$A^{1/2} = V \Lambda^{\frac{1}{2}} V^T \qquad \Lambda^{\frac{1}{2}} = \begin{pmatrix} \lambda_1^{\frac{1}{2}} & & 0 \\ & \ddots & \\ 0 & & \lambda_p^{\frac{1}{2}} \end{pmatrix}$$

$$\begin{aligned} A^{\frac{1}{2}} A^{\frac{1}{2}} &= V \Lambda^{\frac{1}{2}} V^T V \Lambda^{\frac{1}{2}} V^T \\ &= V \Lambda^{\frac{1}{2}} I \Lambda^{\frac{1}{2}} V^T \\ &= V \Lambda V^T = A \end{aligned}$$

- If all of A's e-vals are non-zero
  then A is called invertible/nonsingular

$$A^{-1} = V\Lambda^{-1}V^T$$

$$A^{-1}A = \cancel{V\Lambda^{-1}V^T V\Lambda V^T} = I$$

- If all of A's e-vals are ~~positive~~.
  then A is both PSD & nonsingular
  and is called positive definite (PD)

What about nonsquare matrices?

Answer: singular value decomp (SVD)

For a rectangular matrix $A \in \mathbb{R}^{p \times q}$

$\sigma \in \mathbb{R}$, $v \in \mathbb{R}^p$, $u \in \mathbb{R}^q$ are a
sing val & sing left- & right-sing vector triple
if $A^T v = \sigma u \quad Au = \sigma v$

Any matrix can SVD'ed:

$$A = \sum_{i=1}^{\min(p,q)} \sigma_i u_i v_i^T \quad \text{s.t.} \quad u_i^T u_j = \begin{cases} 1 & i=j \\ 0 & o/w \end{cases} \quad v_i^T v_j = \begin{cases} 1 & i=j \\ 0 & o/w \end{cases}$$

$$= U \Sigma V^T$$

$$U \in \mathbb{R}^{p \times \min(p,q)} \quad V \in \mathbb{R}^{q \times \min(p,q)}$$

$$\Sigma \in \mathbb{R}^{\min(p,q) \times \min(p,q)}$$

$$U^T U = I$$

$$\Sigma = \begin{pmatrix} \ddots & & 0 \\ & \ddots & \\ 0 & & \sigma_{\min(p,q)} \end{pmatrix} \qquad V^T V = I$$

Note:
$$A^T A = V \Sigma \cancel{U^T U} \Sigma V^T = V \Sigma^2 V^T$$
$$A A^T = U \Sigma \cancel{V^T V} \Sigma U^T = U \Sigma^2 U^T$$

Principal Component Analysis (PCA)

Have many hi-dim observation $X \in \mathbb{R}^{n \times p}$

Want to represent these in fewer dims $q < p$ in the most faithful way.

I.e. want to find:

- Vectors $A_1, \cdots, A_q \in \mathbb{R}^p$ i.e. $A = (A_1 \cdots A_q) \in \mathbb{R}^{p \times q}$
- loadings $z_i \in \mathbb{R}^q$ for each x_i

i.e. $Z \in \mathbb{R}^{n \times q}$

such that the approximation

$$\hat{x}_i = \sum_{j=1}^{q} z_{ij} A_j \in \operatorname{span}(\{A_1, \cdots, A_q\})$$

is as close as possible to x_i (simultaneously across $i=1,\cdots,n$)

We may as well let $A_1, \cdots, A_q$ be orthonormal

$$A^T A = I$$

$$\min_{\substack{Z \in \mathbb{R}^{n \times q} \\ p \times q}} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|_2^2 = \sum_{i=1}^{n} \|x_i - A z_i\|_2^2$$

$$= \|X - Z A^T\|^2$$

$A \in \mathbb{R}^{p \times q}$

$A^T A = I$

$\| X - \ldots \|_F$

diff by z_i

$\| M \|_F^2 = \sum_{ij} M_{ij}^2$

$= \| vec(M) \|_2^2$

get

$= A^T (x_i - A z_i) = 0$

$\Rightarrow z_i = A^T x_i$

$$\min_{\substack{A^T A = I \\ A \in \mathbb{R}^{p \times q}}} \overbrace{\| X - X A A^T \|_F^2}^{\sum_{i=1}^n \| x_i - A A^T x_i \|_2^2}$$

Soln: $A = V_q$ ← first q cols of right-sing vecs V from the SVD $A = U \Sigma V^T$

PCA for noncentered data:

PCA: encoders $z = e(x) = A^T x$

decoders $\hat{x} = d(z) = A z$

PCA find e, d s.t. $x \approx d(e(x))$