Lecture 8 - linear & logistic regression

Friday, September 25, 2020 13:23

today: finish linear regression
logistic regression

back to least squares

$$\hat{w}_{MLE} = \arg\min_{w \in \mathbb{R}^d} \|y - Xw\|^2$$

algebra: want ∇_w set to 0

$$\frac{\partial}{\partial w}\left[(y-Xw)^T(y-Xw) \right] \overset{\text{want}}{=} 0$$

$$\frac{\partial}{\partial w}\left[\|y\|^2 - 2y^TXw + w^TX^TXw \right] = 0$$

$$0 - 2X^Ty + 2X^TXw = 0$$

$$\Rightarrow \boxed{(X^TX)w^* = X^Ty}$$ "normal equation"

vector

$$\nabla_w (w^TAw) = (A + A^T)w$$

a) if X^TX ($d \times d$) is invertible, then have unique soln

$$\hat{w}_{MLE} = (X^TX)^{-1}X^Ty$$

X ($n \times d$) $\Rightarrow \text{rank}(X) \leq \min\{n, d\}$

$\text{rank}(X^TX) = \text{rank}(X) \leq \min\{n, d\}$

X^TX is invertible $\Rightarrow \boxed{n \geq d}$

prediction on training set $\hat{y} = X\hat{w} = \underbrace{X(X^TX)^{-1}X^T}_{\hat{w}_{MLE}}y$

projection matrix on column space of X (recall geometric perspective)

if $n < d$ (i.e. high dimension or low data regime) then X^TX is not invertible

b) what if X^TX is not invertible?
→ there is no unique soln

any $\hat{w}$ s.t. $(X^TX)\hat{w} = X^Ty$ is a MLE estimate

could choose $\hat{w} = \arg\min_{w:\ (X^TX)w = X^Ty} \|w\|_2 \Rightarrow X^+y$ ← Moore-Penrose pseudo-inverse (see Wikipedia)

$X^+ = (X^TX)^{-1}X^T$

when X is full rank

$X = U \Sigma V^T \quad X^+ = V \Sigma^+ U^T$

$n \times d \qquad d \times n$

$\Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_d \end{pmatrix}$

problem: pseudo-inverse is not numerically stable

instead it is better to regularize to get similar effect

$\hat{w}_{MAP}(\lambda) \xrightarrow{\lambda \to 0} \hat{w}_{pseudo\text{-}inverse}$

regularization (can be motivated from MAP pt. of view)

Suppose we put a prior $p(w) = N(w \mid 0, \sigma^2 I)$ (I: $d \times d$ identity matrix)

$\lambda \leftarrow$ "precision" parameter

log posterior: $\log p(w \mid data) = \log p(y_{1:n} \mid X, w) + \log p(w) + cst.$

$= \frac{-1}{2\sigma^2} \|y - Xw\|_2^2 + f(\sigma^2) \frac{-\lambda}{2\sigma^2} \|w\|^2 + cst.$

MAP here $\hat{w}_{MAP} = \arg\min_w \frac{1}{2} \|y - Xw\|^2 + \frac{\lambda}{2} \|w\|_2^2$ "ridge regression"

↳ same as "regularized" ERM

empirical risk minimization

$\frac{1}{n} \sum_{i=1}^{n} \underbrace{l(y_i, h_w(x_i))}_{\text{squared loss}} + \frac{\lambda}{2n} \|w\|^2$

empirical error — regularization

this is objective is strongly convex in w $\Rightarrow$ a unique soln

$f(\cdot)$ is λ-strongly convex $\Leftrightarrow$ $f(\cdot) - \frac{\lambda}{2}\|w\|^2$ is convex in $(\cdot)$

$\nabla_w = 0 \Rightarrow (X^T X + \lambda I) w = X^T y$

always invertible for $\lambda > 0$

$\hat{w}_{MAP} = (X^T X + \lambda I)^{-1} X^T y$ — no problem for $d > n$

ridge regression

14h39

good practice: to either standardize features ie. make each feature zero mean & unit empirical variance

or

normalize ← make x_i unit norm $\|x_i\|_2 = 1$

or

scale features to $[0,1]$ or $[-1,1]$

Logistic regression

setup: binary classification $Y = \{0,1\}$ $X \in \mathbb{R}^d$

generative model motivation:

suppose only assumption is there exists a pdf (densities) in $\mathbb{R}^d$ for the class conditionals:

$$p(x|Y=1) \;\&\; p(x|Y=0)$$

$$P(Y=1|X=x) = \frac{p(Y=1, X=x)}{\{p(Y=1, X=x) + p(Y=0, X=x)\}\; p(X=x)}$$

$$= \frac{1}{1 + \frac{p(Y=0, X=x)}{p(Y=1, X=x)}} = \frac{1}{1+\exp(-f(x))}$$

where

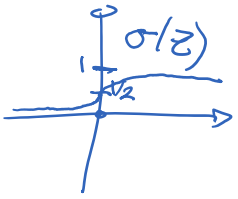
$$f(x) \triangleq \log \frac{p(X=x|Y=1)}{p(X=x|Y=0)} + \log \frac{P(Y=1)}{P(Y=0)}$$

"log odds" — class conditional ratio — prior odd ratio

in general, $P(Y=1|X=x) = \sigma(f(x))$

where $\sigma(z) \triangleq \frac{1}{1+\exp(-z)}$

"sigmoid function"



some properties of $\sigma(z)$:

$$\sigma(-z) = 1 - \sigma(z) \qquad [\sigma(z) + \sigma(-z) = 1]$$

$$\frac{d\sigma(z)}{dz} = \sigma(z)(1-\sigma(z)) = \sigma(z)\sigma(-z)$$

(*) to motivate linear logistic regression, consider class conditionals in the exponential family

$$p(x|\eta) \triangleq h(x)\exp(\eta^T T(x) - A(\eta))$$

- η: "canonical parameters"
- $T(x)$: "sufficient statistics"
- $A(\eta)$: scalar function; log partition function; normalizer
- $h(x)$: these specify the "flat" exponential family that we are considering

Gaussian: $\log p(x|\mu,\sigma^2) = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}$

$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

$-\left[\frac{x^2}{2\sigma^2} - x\frac{\mu}{\sigma^2} + \frac{\mu^2}{2\sigma^2}\right]$

let $T(x) = \begin{bmatrix} -\frac{x^2}{2} \\ x \end{bmatrix}$

$\eta(\mu, \sigma^2) = \begin{bmatrix} 1/\sigma^2 \\ \mu/\sigma^2 \end{bmatrix}$

$A(\eta) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{\mu^2}{2\sigma^2}$

$p(x|Y=1) = p(x|\eta_1)$

$p(x|Y=0) = p(x|\eta_0)$

log odds $f(x) = \log \underbrace{\frac{p(x|Y=1)}{p(x|Y=0)}}_{p(x|\eta_1)/p(x|\eta_0)} + \log \underbrace{\frac{p(Y=1)}{p(Y=0)}}_{\pi/(1-\pi)}$ ($\triangleq \pi$)

$= (\eta_1 - \eta_0)^T T(x) + A(\eta_0) - A(\eta_1) + \log\frac{\pi}{1-\pi}$

$\triangleq w^T \varphi(x)$

where $w = \begin{pmatrix} \eta_1 - \eta_0 \\ A(\eta_0) - A(\eta_1) + \log\frac{\pi}{1-\pi} \end{pmatrix}$ $\varphi(x) = \begin{pmatrix} T(x) \\ 1 \end{pmatrix}$

get logistic regression model $\boxed{P_w(Y=1|X=x) = \sigma(w^T \varphi(x))}$

"feature map"

Exercise to reader:

try argument above with $p(x|y) = N(x|\mu_y, \Sigma_y)$

if $\Sigma_0 = \Sigma_1$, then $\varphi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}$

otherwise $\varphi(x) = \begin{pmatrix} xx^T \\ x \\ 1 \end{pmatrix}$ (optional)