Lecture 7 - scribbles

Tuesday, September 26, 2017
14:35

today: · finish linear regression
· logistic regression

Linear regression:

$$\hat{w}_{MLE} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \|\vec{y} - Xw\|^2$$

"vector" convention

algebra: want $\vec{\nabla}_w = 0$

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

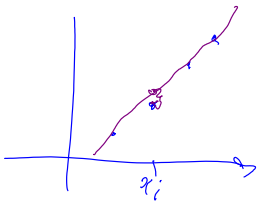
$$\frac{\partial}{\partial w}\left(\|y\|^2 - 2y^TXw + w^TX^TXw\right) = 0$$

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

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

constant

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



if X^TX is invertible, then have unique solution

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

$\underset{d\times n}{X^T}\underset{n\times d}{X} \to d \times d$ matrix

$\operatorname{rank}(\underset{n\times d}{X}) \le \min\{n, d\}$

prediction on training set: $\hat{y} = X\hat{w} = \underbrace{X(X^TX)^{-1}X^T}_{\text{projection operator on column space of } X}\, y$

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

(*) what if X^TX is not invertible?

⊛ what if X^TX is not invertible?

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

could choose $\hat{w} = \underset{w:\, X^TXw = X^Ty}{\operatorname{argmin}} \|w\| = X^\dagger y$ (Moore–Penrose pseudo-inverse)

problem: pseudo-inverse is not numerically stable

instead it is better to regularize to get similar effect

regularization (from MAP point view):

suppose we put prior $p(w) = N(w|0, \sigma^2 I)$ ← $d \times d$ identity; λ "precision" parameter ← multivariate $N(\vec{\mu}, \Sigma)$, $d \times d$

log posterior: $\log p(w|\underset{X,\vec{y}}{\text{data}}) = \log p(y_{1:n}|x_{1:n}, w) + \log p(w) + \text{cst.}$

($\log p(y_{1:n}|x_{1:n}, w)$: conditional likelihood; $\log p(w)$: $\downarrow$)

$= -\frac{\|\vec{y} - X\vec{w}\|^2}{2\sigma^2} + \text{cst.}(\sigma^2) - \frac{\lambda\|w\|^2}{2\sigma^2} + \text{cst.}$

MAP here

$$\hat{w}_{MAP} = \underset{w}{\operatorname{argmin}} \frac{1}{2}\|\vec{y} - X\vec{w}\|^2 + \frac{\lambda\|w\|^2}{2}$$

"ridge regression"

same as "regularized ERM" $\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_w(x_i)) + \frac{\lambda}{2}\|w\|^2$

squared loss; regularization

empirical error

is "strongly convex" in w

$f(\cdot)$ is strongly convex $\Leftrightarrow f(\cdot) - \frac{\lambda\|\cdot\|^2}{2}$ is convex for some $\lambda > 0$

$\nabla_w = 0 \Rightarrow (X^TX + \lambda I)w = X^Ty$

$X^TX + \lambda I$: always invertible

$\Downarrow$

unique solution

$\nabla w = 0 \Rightarrow \underbrace{(X^TX + \lambda I)}_{\text{always invertible (for } \lambda > 0)} w = X^T y$ ⇓ unique solution

$$\hat{w}_{MAP} = (X^TX + \lambda I)^{-1} X^T y$$

ridge or regression

one comment:

good practice to either standardize features i.e. make each feature zero mean and unit empirical variance

or

normalize features — make x_i unit norm ($\|x_i\| = 1$) or scale features to $[0,1]$ or $[-1,1]$

Logistic regression:

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

motivation: suppose only assumption is $\exists$ pdfs (densities) in $\mathbb{R}^d$

$p(x|Y=1)$ and $p(x|Y=0)$ ← "class conditionals"

$$P(Y=1|X=x) = \frac{P(Y=1, X=x)}{P(Y=1,X=x) + P(Y=0,X=x)} \quad \left\} P(X=x)\right.$$

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

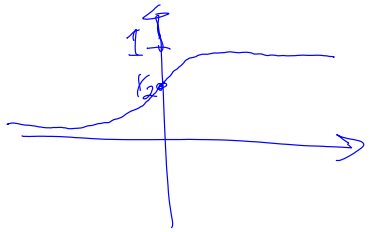
where $f(x) \triangleq \log \underbrace{\frac{P(X=x|Y=1)}{P(X=x|Y=0)}}_{\text{class-conditional ratio}} + \log \underbrace{\frac{P(Y=1)}{P(Y=0)}}_{\text{prior odds ratio}}$

"log odds"

and thus in general $P(Y=1 \mid X=x) = \sigma(f(x))$

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

"sigmoid function"



some properties of $\sigma(z)$:

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

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

To motivate linear logistic regression, consider class conditionals in the exponential family

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

η: "canonical parameter"; $h(x)$, $T(x)$: these specify the family; $A(\eta)$: scalar function = log partition function; $T(x)$: "sufficient statistics"; linear in η

$$\text{log odds } f(x) = \log \frac{p(x \mid Y=1)}{p(x \mid Y=0)} + \log \frac{p(Y=1)}{p(Y=0)}$$

$p(x \mid Y=1) = p(x \mid \eta_1)$, $p(x \mid Y=0) = p(x \mid \eta_0)$, $p(Y=1) = \pi$, $p(Y=0) = 1-\pi$

$\eta_0 \rightarrow$ parameter for class 0

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

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

"feature map"

def logistic regression model

get logistic regression model

$$p_w(Y=1 \mid x=x) = \sigma(w^T \phi(x))$$

("feature map")

logic regression model: $Y = \{0, 1\}$

$$p(y=1 \mid x) = \sigma(w^T x)$$

$$p(y=0 \mid x) = 1 - \sigma(w^T x) = \sigma(-w^T x)$$

[if $Y = \{\pm 1\}$ encode $p(y \mid x) = \sigma(y w^T x)$]

$Y \mid X=x$ is a Bernoulli$(\sigma(w^T x))$

$$p(y \mid x) = \sigma(w^T x)^y \sigma(-w^T x)^{1-y}$$

given $(x_i, y_i)_{i=1}^n$, maximum conditional log-likelihood:

$$\ell(w) = \sum_{i=1}^n \log p(y_i \mid x_i; w) = \sum_{i=1}^n [y_i \log \sigma(w^T x_i) + (1-y_i) \log \sigma(-w^T x_i)]$$

$$\sigma'(z) = \sigma(z)\sigma(-z)$$

$$\nabla_w \sigma(w^T x_i) = x_i [\sigma(w^T x_i)\sigma(-w^T x_i)] \qquad \text{let } v_i \triangleq w^T x_i$$

$$\nabla \ell(w) = \sum_{i=1}^n x_i \left[y_i \frac{\sigma(v_i)\sigma(-v_i)}{\sigma(v_i)} - (1-y_i) \frac{\sigma(-v_i)\sigma(v_i)}{\sigma(-v_i)} \right]$$

$$[y_i (\underbrace{\sigma(-v_i) + \sigma(v_i)}_{1}) - \sigma(v_i)]$$

$$\boxed{\nabla \ell(w) = \sum_{i=1}^n x_i [y_i - \sigma(w^T x_i)]}$$

need to use numerical optimization

solving for $\nabla \ell(w) = 0 \Rightarrow$ need to solve a transcendental eq.

solving for $\nabla \ell(w) = 0$ $\Rightarrow$ need to solve a transcendental eq.

because $\frac{1}{1+\exp(-w^T x_i)}(\dots) = 0$

(need to solve for this)

(contrast to least square (linear in w)

$$\nabla \ell(w) = \sum_i x_i [y_i - w^T x_i]$$