Lecture 8 - scribbles

Tuesday, October 1, 2019 14:28

today : • logistic regression
• numerical optimization (IRLS)

comment about linear regression:

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 $\|x_i\|_2 = 1$

or

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

Logistic regression

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

generative model motivation:

suppose only assumption is $\exists$ a pdf (densities) in $\mathbb{R}^d$

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

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

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

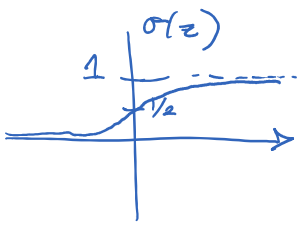
where $f(x) \triangleq \log \frac{P(X=x|Y=1)}{P(X=x|Y=0)}$ (class conditional ratio) $+ \log \frac{P(Y=1)}{P(Y=0)}$ (prior odds ratio)

"log odds"

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

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

"sigmoid function"



some properties of $\sigma(z)$:

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

$$d\sigma(z) = \sigma(z)(1-\sigma(z)) = \sigma(z)\sigma(-z)$$

dz

(*) 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 parameter"
- h, T: these specify the "flat" family
- $T(x)$: "sufficient statistics"
- A: scalar fct: "log partition fct." / normalizer

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

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

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

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

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

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

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

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

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

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

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

get logistic regression model

$$P_w(Y=1|X=x) = \sigma(w^T \varphi(x))$$

φ: "feature map"

15h34

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 $\varphi(x) = \begin{pmatrix} xx^T \\ x \\ 1 \end{pmatrix}$ (if you want)

logistic regression Model:

$p(y=1|x) = \sigma(w^T x)$ $\quad Y = \{0,1\}$ $\quad$ [if $Y = \{-1, 1\}$ encode $p(y|x) = \sigma(y w^T x)$]

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

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

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

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

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

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

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

$$= \sum_{i=1}^n x_i [y_i [\sigma(-v_i) + \sigma(v_i)] - \sigma(v_i)]$$

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

need to use numerical methods

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

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

contrast to linear regression: $\nabla \ell(w) = \sum_{i=1}^n x_i [y_i - w^T x_i]$ (linear in w)

numerical optimization

want to minimize $f(w)$ (unconstrained)

s.t. $w \in \mathbb{R}^d$

1) **gradient descent** (1st order method)

start at w_0

iterate: $w_{t+1} = w_t - \underbrace{\gamma_t}_{\text{step-size}} \nabla f(w_t)$

stopping criterion: $\|\nabla f(w_t)\| < \delta$ e.g. $\delta = 10^{-6}$

note: f μ-strongly-convex $\Leftrightarrow$ $f(x) - \frac{\mu}{2}\|x\|^2$ is convex

$\Rightarrow \|\nabla f(w_t)\|$ small $\Rightarrow f(w_t) - f(w^*)$ is also small $\cdot \frac{1}{\mu}$

step-size rules:

a) constant step-size $\gamma_t = \frac{1}{L}$ ← Lipschitz continuity constant of ∇f

$\|\nabla f(w) - \nabla f(w')\| \leq L \|w - w'\|$

b) decreasing step-size rule: $\gamma_t = \frac{c}{t}$ ← constant

usually want: $\sum_t \gamma_t = \infty$ $\sum_t \gamma_t^2 < \infty$

(more used for stochastic optimization: $f(w) = \mathbb{E}_\xi g(w, \xi)$)

c) choose γ_t by "line search": $\min_{\gamma \in \mathbb{R}} f(w_t + \gamma d_t)$

d_t: direction for update e.g. $-\nabla f(w_t)$

↓ costly in general

instead do approximate search e.g. Armijo line search (see Boyd's book)

Newton's method (2nd order method)

motivation: minimizing a quadratic approximation:

Taylor expansion: $f(w) = f(w_t) + \nabla f(w_t)^\top (w - w_t) + \frac{1}{2}(w - w_t)^\top H(w_t)(w - w_t) + O(\|w - w_t\|^3)$

Hessian $[H(w_t)]_{ij} = \frac{\partial^2 f(w_t)}{\partial w_i \partial w_j}$

$O(\|w - w_t\|^3)$: Taylor's remainder

$$f(w_t) = Q_t(w) + O(\|w - w_t\|^3)$$

↑ quadratic model approximation

$w_{t+1} \rightsquigarrow$ minimizing $Q_t(w)$

$$\nabla_w Q_t(w) = 0 \qquad \nabla f(w_t) + H(w_t)(w - w_t) = 0$$

$$\Rightarrow \quad w - w_t = -H^{-1}(w_t)\nabla f(w_t)$$

$$\boxed{w_{t+1} = w_t - H^{-1}(w_t)\nabla f(w_t)}$$ Newton's update

↓ inverse Hessian $\rightsquigarrow O(d^3)$ time to compute in general and $O(d^2)$ space

note: in practice and for assignment:

numpy.linalg.lstsq to find $H^{-1}b$ as solution to $\min_w \|Hw - b\|^2$

↑ numerically more stable than