Lecture 9 - scribbles

Friday, October 4, 2019 14:01

today: • finish Newton + IRLS
• Fisher LDA + math tricks

Newton's continuation:

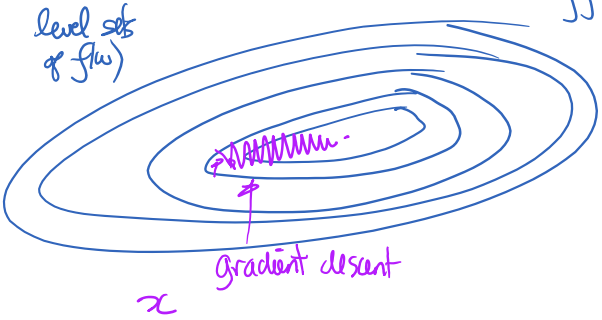
damped Newton: you add a step-size to stabilize Newton's method

$$w_{t+1} = w_t - \gamma H^{-1}(w_t) \nabla f(w_t)$$

γ: step-size

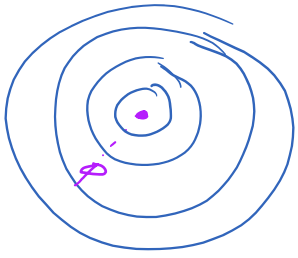
why Newton's method?

- much faster convergence in # of iterations vs. gradient descent
- affine covariant → method is invariant to rescaling of variables



Newton is using Hessian to make f "well-conditioned"

$z = H^{1/2} x$



"quadratic form"

$$\frac{1}{2} x^T H x = c$$

$$H = P^T \Sigma P$$

$$\frac{1}{2} x^T P^T \Sigma P x = c$$

Σ: diagonal

H is symmetric and PSD

$$z = H^{1/2} x = \Sigma^{1/2} P x$$

$$\frac{1}{2} z^T z = c$$

exercise to reader:

$$z_{t+1} = z_t - \gamma \nabla f(z_t)$$

$$\Updownarrow$$

$$x_{t+1} = x_t - \gamma H^{-1} \nabla f(x_t)$$

Newton's method for logistic regression: IRLS

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

$$H(\ell(w)) = -\sum_{i=1}^{n} x_i x_i^T \sigma(w^T x_i) \sigma(-w^T x_i)$$

$$v^T H v = -\sum (v^T x_i)(x_i^T v) \sigma(-) \sigma( )$$

$(v^T x_i)^2$; $\sigma(-)\sigma( ) \geq 0$

$v^T H v \leq 0 \quad \forall v \in \mathbb{R}^d$

ie. $H \preceq 0$

$v^T \nabla^2 \ell(w) v = -\sum_i \underbrace{(v^T x_i)^2}_{\geq 0} \underbrace{\mu_i(1-\mu_i)}_{\geq 0}$ ≤ 0 even.

ie. $H \leq 0$

i.e. concave fct.

$\Downarrow$

Newton is max. instead of min.

notation: recall $X_{n \times d} = \begin{pmatrix} \vdots \\ -x_i^T- \\ \vdots \end{pmatrix}$

let $\mu_i \triangleq \sigma(w^T x_i) \in]0,1[$

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

Hessian $= -\sum_i x_i x_i^T \mu_i(1-\mu_i) = -X^T D(w) X$ where $D_{ii} = \mu_i(1-\mu_i)$ (depends on w)

Newton's update: $w_{t+1} = w_t - (-X^T D_t X)^{-1} X^T(y-\mu_t)$

$w_{t+1} = (X^T D_t X)^{-1}\left[(X^T D_t X) w_t + X^T(y-\mu_t)\right]$

$\boxed{w_{t+1} = (X^T D_t X)^{-1}[X^T D_t z_t]}$

where $\boxed{z_t \triangleq X w_t + D_t^{-1}(y-\mu_t)}$

this is a solution to "weighted least square problem"

$\min_w \| D^{1/2}(z_t - Xw) \|^2$ (D: weights, z_t: new target)

$\sum_i (z_i - x_i^T w)^2$ ($d_{ii} \triangleq$)

compare with Gaussian noise model for least sq. $\sum_i \frac{(y_i - x_i^T w)^2}{2\sigma^2}$

Newton's method for logistic regression

= Iterative reweighted least square (IRLS)

note: $x = A^{-1}b$ $Ax = b$ $\min_x \|Ax-b\|^2$

Big data logistic regression: suppose $f(w) = \frac{1}{n}\sum_{i=1}^n f_i(w)$

- big d $\Rightarrow$ cannot do $O(d^2)$ or $O(d^3)$ operations $\Rightarrow$ first order methods
- if n is huge, you cannot do batch methods $\nabla f(w) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w)$ (∇f_i: gradient of one fct.) $O(nd)$ time
  ($\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w)$: batch gradient)

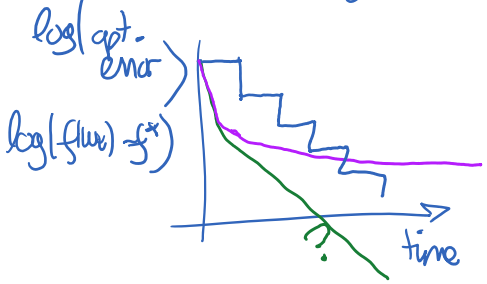
  instead you "incremental gradient methods"

  eg. stochastic gradient descent (SGD): $w_{t+1} = w_t - \gamma_t \nabla f_{i_t}(w_t)$ $O(d)$ time

  where i_t is picked uniformly at random

SGD → cheap updates, but slower convergence per iteration

batch gradient → expensive updates, but faster " "



variance reduced methods

SAG: stochastic averaged gradient [2012]

G.D. $w_{t+1} = w_t - \gamma_t \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_t)$

SAG: $w_{t+1} = w_t - \gamma_t \frac{1}{n}\sum_{i=1}^{n} v_i^t$ (memory)

where $v_i = \nabla f_i(w_{old,i})$

at each t, update only one $v_{i_t} \triangleq \nabla f_{i_t}(w_t)$

SAGA: $w_{t+1} = w_t - \gamma\left(\nabla f_{i_t}(w_t) + \frac{1}{n}\sum_{j} v_j - v_{i_t}\right)$ [2014]

($\frac{1}{n}\sum_{j} v_j - v_{i_t}$: variance reduction correction)

(default method for log. regression in scikit-learn)

15h12

generative model for classification: (Fisher) linear discriminant analysis

FLD (LDA)

for classification $Y \in \{0,1\}$, $X \in \mathbb{R}^d$

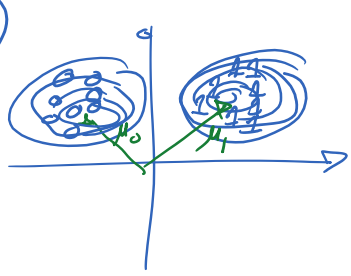
generative approach: $p(x,y;\theta) = \overbrace{p(x|y;\theta)}^{\text{class conditionals}} p(y;\theta)$

vs.

conditional approach $p(y|x;\theta)$

for Fisher model: we assume $p(x|y;\theta) = N(x|\mu_y, \Sigma)$ (Σ: shared for both classes)

for Fisher model: we assume $p(x|y;\theta) = N(x|\mu_y, \Sigma)$ ← classes



$\theta = (\mu_0, \mu_1, \Sigma, \pi)$

μ_0, μ_1: mean for class 0; Σ: shared; π: $p(y=1)$

can then show that $p(y|x;\theta) = \sigma(w^T x)$

where w is a fct. of $(\mu_0, \mu_1, \Sigma, \pi)$

[note: if you use Σ_0 & Σ_1, get "quadratic discriminant analysis" (QDA)

i.e. $\sigma(w^T \varphi(x))$ where $\varphi(x)$ is quadratic fct. of x

↝ see hwk 2]

(*) gen. approach: do joint MLE to estimate

$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_i \log p(x_i, y_i; \theta)$

vs. log. reg. which MCL

sidenote: MLE for multivariate Gaussian

$X_i \sim N(\mu, \Sigma)$ $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$, Σ is symmetric, $\Sigma \succ 0$

$\Sigma = \mathbb{E}[(x-\mu)(x-\mu)^T]$

$\Sigma^T = \mathbb{E}[\ \ \ ] = \Sigma$

$$p(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$\theta = (\mu, \Sigma)$

$\operatorname{tr}(AB) = \operatorname{tr}(BA)$

$\operatorname{tr}\left((x-\mu)^T \Sigma^{-1} (x-\mu)\right)$

$\operatorname{tr}\left(\Sigma^{-1} (x-\mu)(x-\mu)^T\right) = \langle \Sigma^{-1}, (x-\mu)(x-\mu)^T \rangle$ ← inner product

$\langle A, B \rangle = \sum_{i,j} A_{ij} B_{ij} = \operatorname{tr}(A^T B)$

log-likelihood: $\sum_{i=1}^n \log p(x_i;\theta) = \text{cst.} - \frac{n}{2} \log|\Sigma| - \frac{n}{2} \left\langle \Sigma^{-1}, \frac{1}{n} \sum_{i=1}^n (x_i-\mu)(x_i-\mu)^T \right\rangle$

$\frac{1}{n} \sum_{i=1}^n (x_i-\mu)(x_i-\mu)^T = \hat{\Sigma}(\mu)$

vector derivative review:

suppose $f: \mathbb{R}^m \to \mathbb{R}^n$

f is differentiable at x iff $\exists$ linear operator $df: \mathbb{R}^m \to \mathbb{R}^n$ $o(\|\Delta\|)$ either dn means is

Suppose $f: \mathbb{R}^m \to \mathbb{R}^n$

f is differentiable at x_0 iff $\exists$ a linear operator $df_{x_0}: \mathbb{R}^m \to \mathbb{R}^n$

s.t. $\forall \Delta \in \mathbb{R}^m \quad f(x_0+\Delta) - f(x_0) = df_{x_0}(\Delta) + o(\|\Delta\|)$

("differential")

$o(\|\Delta\|)$ "little oh" means is some fct. $h(\|\Delta\|)$ s.t. $\lim_{\|\Delta\|\to 0} \frac{h(\|\Delta\|)}{\|\Delta\|} \to 0$

df_{x_0} is linear means $df_{x_0}(\Delta_1 + b\Delta_2) = df_{x_0}(\Delta_1) + b\, df_{x_0}(\Delta_2)$

can represent as a $n \times m$ matrix called the Jacobian matrix

standard representation $(df_{x_0})_{ij} = \frac{\partial f_i}{\partial x_j}$ (i^{th} comp. of f, j^{th} component of x)

1) this gives a way to get df_{x_0} for anything (matrix, tensor, ∞-dim fct.)

2) be careful with dimension: $f: \mathbb{R}^m \to \mathbb{R}$

df_{x_0} is a row vector $(1 \times m)$ $\quad df_{x_0} = (\nabla f(x_0))^T$

Chain rule: Suppose $f: \mathbb{R}^m \to \mathbb{R}^n$, $g: \mathbb{R}^n \to \mathbb{R}^q$

$$d(g \circ f)_{x_0} = dg_{f(x_0)} \circ df_{x_0}$$

(product of Jacobians)

e.g. $f(x) = x - \mu \qquad df_{x_0} = I$

$g(x) = x^T A x \qquad dg_{x_0} = x^T(A + A^T)$

$g(f(x)) = (x-\mu)^T A (x-\mu) \qquad d(g \circ f)_{x_0} = dg_{f(x_0)} \circ df_{x_0}$

$= (x-\mu)^T (A + A^T) \cdot I$

for Gaussian: $-\frac{1}{2} \sum_i (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$

$\nabla_\mu \quad -\frac{1}{2} \sum_i 2\Sigma^{-1}(x_i - \mu) \overset{\text{want}}{=} 0$

$$\Rightarrow \hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

some references:

- for convex optimization: Boyd & Vandenberghe's book

- DL book -- chapter 4.3 on gradient-based optimization

- for matrix calculus:
  Matrix Differential Calculus with Applications in Statistics and Econometrics, Heinz Neudecker and Jan R. Magnus -- free online