Tutorials for Midterm

Thursday, November 13, 2014 3:04 PM

Note: Due to the nature of the tutorials, things are a bit disjointed.
I have labeled most of the stuff.
Good luck everyone! Exept Issam, you don't need luck =)
-Scott

Some Matrix Properties

$$f(x) = x^T A x + bx + c$$

$$\nabla f(x) = Ax + b$$

ie: $\frac{1}{2}(A + A^T) + b$

symetric

$$f(x) = \|Ax - b\|^2$$

$$= \tfrac{1}{2}\left(b^T b - 2x^T A^T b + x^T A^T A x\right)$$

$$\nabla f(x) = 0 - A^T b + A^T A x$$

$$= A^T(Ax - b)$$

$$\sum_i v_i x_i = v^T x = x^T v$$

$v = A^T B C Y w$

$(n \cdots\cdots x_1)$

$$\sum_i x_i \sum_j x_j a_{ij} = x^T A x$$

$$\sum_i x_i v_i = x^T v$$

$v_j = x_j a_{ij}$

$$Ax = \begin{bmatrix} \sum_j x_j a_{1j} \\ \vdots \\ \sum_j x_j a_{nj} \end{bmatrix}$$

$f = X^T A X$

$\nabla f = AX$

$\nabla^2 f = A$

✓

$$\min_w f(w) + \frac{\lambda}{2}\|w\|^2$$

$$w^{t+1} = w^t - \alpha\left(\nabla f(w^t) + \lambda w^t\right)$$

Grad descent

$$w^{t+1} = w^t - \alpha \nabla g(w^t)$$

$$\underset{d \text{ by } 1}{w^t} = \underset{1 \text{ by } 1}{\beta^t}\, \underset{d \text{ by } 1}{v^t}$$

ex

$$w^t - \alpha\lambda w^t - \alpha \nabla f_i(w^t) = \underset{1 \text{ by } 1}{(1 - \alpha\lambda)}\underset{d \text{ by } 1}{w^t} - \alpha f_i(x^t)$$

$$w^{t+1} \overset{\text{proportional}}{=} (1 - \alpha\lambda) w^t \qquad O(d)$$

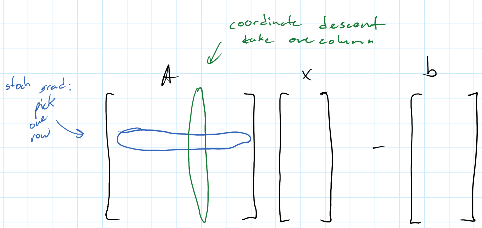
$$\beta^{t+1} v^{t+1} = (1 - \alpha\lambda)\beta^t v^t$$

$$= \left[(1 - \alpha\lambda)\beta^t\right] v^t$$

$$v^{t+1} = v^t$$

$$\beta^{t+1} = (1 - \alpha\lambda)\beta^t \qquad O(1)$$

Stoch Grad vs Coord Desc.



On Beta and Bernoulli

$x_i \sim Ber(\theta)$

$\theta \sim Beta(a, b)$

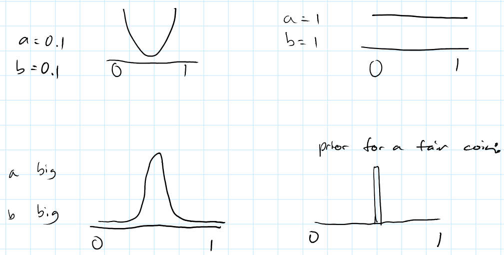
$p(x_i) = \theta^{I[x_i=1]}(1-\theta)^{I[x_i=0]}$ likelihood

$p(\theta | a, b) = \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}$

prior

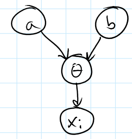
$\propto \theta^{a-1}(1-\theta)^{b-1}$

Beta function



$p(\theta | x_i, a, b) \propto p(x_i | \theta, a, b)\, p(\theta | a, b)$

$= p(x_i | \theta)\, p(\theta | a, b)$



if I have θ, a, b have no effect

$\propto \left[\theta^{I(x_i=1)}(1-\theta)^{I(x_i=0)}\right]\left[\theta^{a-1}(1-\theta)^{b-1}\right]$

$\propto \left[\theta^{I(x_i=1)+a-1}(1-\theta)^{I(x_i=0)+b-1}\right]$

$\theta | x_i, a, b \sim Beta(I[x_i=1]+a,\ I[x_i=0]+b)$

new prior

$\int p(\theta | a, b)\, d\theta = 1$

$\int \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} = 1$

$\frac{1}{B(a,b)} \int \theta^{a-1}(1-\theta)^{b-1} = 1$

$\int \theta^{a-1}(1-\theta)^{b-1} = B(a,b)$

$\text{mean}(Beta) = \frac{a}{a+b}$

Linear Constraints to non-smooth problem

step1. write in maxes

$\sum_{i=1}^{N} |w^T x_i - y_i| + \lambda \|w\|_\infty$

$|x| = \max\{x, -x\}$

$\sum_{i=1}^{N} \max\{w^T x_i - y_i,\ y_i - w^T x_i\} + \lambda \max_j \{|w_j|\}$

$\Downarrow$

$\lambda \max_j \{\max\{w_j, -w_j\}\}$

step 2
upper bound max with new variables

$\sum_{i=1}^{N} v_i \quad \text{s.t.} \quad v_i \geq \max\{w^T x_i - y_i,\ y_i - w^T x_i\}$

$\min_x f(x) \iff \min_{x,v} v \quad \text{s.t. } v \geq x$

(at soln $v = f(x)$)

$\max_j \{\max\{w_j, -w_j\}\} \Rightarrow r \qquad \text{s.t. } r \geq \max_j \{w_j, -w_j\}$

step 3

$v \geq \max_i \{x_i\} \Rightarrow v \geq x_i \quad \forall i$

$v \geq \max(a, b)$

$\Rightarrow v \geq a,$
$v \geq b$

$$\min_{w, v, r} \sum_{i=1}^{N} v_i + \lambda r \quad \text{s.t.}$$

$v_i \geq w^T x_i - y_i$
$v_i \geq y_i - w^T x_i$

$r \geq w_j$
$r \geq -w_j$

LP: $\min_x c^T x$ s.t. $Ax \leq b$

QP: $\min_x \frac{1}{2} x^T H x + c^T x$ s.t. $Ax \leq b$

prox: $\min_x f(x) + r(x)$ ie. $r(x)$ non continuous

$$x^{t+1} = \text{prox}_{\alpha r(\cdot)} [x^t - \alpha_t \nabla f(x^t)]$$

$$\text{prox}_{\alpha r(\cdot)} [y] = \underset{x}{\text{argmin}} \frac{1}{2} \|x - y\|^2 + \alpha r(y)$$

proj

$$r(x) = \begin{cases} 0 & \text{if } Ax \leq b \\ \infty & \text{if } Ax > b \end{cases}$$

QP: $\min_x \underbrace{\frac{1}{2} x^T H x + c^T x}_{f(x)} + r(x)$

On time to error

$\|x^t - x^*\| = O(1/t)$ how big t, such that $x^t - x^* \leq \epsilon$

$\leq c \frac{1}{t}$

$\leq \epsilon$

$O(1/t^2) \leq 1/t^2$
$O(p^t) \leq c p^t$

$\frac{c}{t} \leq \epsilon \Rightarrow \frac{c}{\epsilon} \leq t \qquad t \geq O(1/\epsilon)$

$t = \Omega(1/\epsilon)$

Dual norms

$f(x) = \|x\|_p$

$$f^*(y) = \begin{cases} 0 & \text{if } \|y\|_q \leq 1 \\ \infty & \text{else} \end{cases}$$

$1 \rightarrow \infty$
$2 \rightarrow 2$

$\frac{1}{p} + \frac{1}{q} = 1$

$f(x) = \|x\|_p^2 \qquad \lambda \|x\|_p^2$

$f^*(y) = \|y\|_q^2 \qquad \frac{1}{\lambda} \|y\|_q^2$

if $g(x) = a f(x)$

then $g^*(y) = a f\left(\frac{y}{a}\right)$

conjugate

$$f^*(y) = \sup_x \{ y^T x - f(x) \}$$

$$f(x) = \exp(x)$$

$$f^*(y) = \sup_x \{ \underbrace{yx - \exp(x)}_{g(x)} \}$$

$$\nabla g(x) = y - \exp(x)$$

at minimum, $\nabla g(x) = 0$

$$0 = y - \exp(x)$$

solve for x

$$x = \log(y) \qquad (\text{for } y > 0)$$

plug x back in

$$f^*(y) = y\log(y) - \exp(\log(y))$$

$$= y\log y - y$$

$$f^*(y) = y(\log(y) - 1) \qquad \text{domain: } y > 0$$

$$f^*(y) = \begin{cases} y(\log(y) - 1) & y > 0 \\ \infty & y \leq 0 \end{cases}$$

"separable" function

$$\min \sum_{i=1}^{d} f_i(x_i)$$ different f to each variable

ex: $f(x) = x_1 + x_2^2 + \exp(-x_3) + \tanh(x_4) \cdots$

$$\min_{x_1 x_2 x_3 x_4 \cdots} f(x)$$ is independent

equivalent:

$$\sum_{i=1}^{d} \min_{x_i} \{ f(x_i) \}$$

- can't have $x_1 = x_2$

EM:

θ^t $\downarrow$ E step

m step: maximize to get new θ^t

$$Q(\theta \mid \theta^t) = \sum_i \sum_h w_{ih} \log p(x, h \mid \theta)$$

Tutorial Part 2

EM :
(will be bonus)

$$Q(\theta|\theta^t) \sum_{i=1}^{N} \sum_{h} w_{ih} \log p(x,h|\theta)$$

$\uparrow \theta^t$

– sum outside log, not inside

Prisoners dilema:

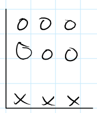
$$p(F|I) = 0.01$$ footprints, given innocent

$$p(\neg F|I) = 1 - p(F|I)$$

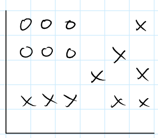
$$p(F|I) \propto p(I|F)\,p(F)$$

$\hookrightarrow 1 - p(\neg I|F)$

Algorithms
–may ask about runtime



decision stumps?



how about now?
–maybe with boosts

(interpret plot)

techniques

← maybe 1 norm?
some other norm

1. $$f(x^{k+1}) \geq f(x^k) - \frac{1}{2\mu}\|\nabla f(x^k)\|^2$$

likely on test!

2. $$f(x^*) \leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2$$

combine inequalities for solution

$$f(x^{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)\left[f(x^k) - f(x^*)\right.$$

$$\|x^{k+1} - x^*\|^2$$
$$= \|(x^k - \alpha_k \nabla f(x^k)) - x^*\|^2$$
$$= \|(x^k - x^*) - \alpha_k \nabla f(x^k)\|^2$$

on test: replace l_2 with l_1

ber with multi

etc...

(small changes to questions you've seen) coding: implement... ex NB

KNN, with ℓ_1 norm

start/end of code similar to assn

i.e. will have a start point

Define stuff on your cheat sheet if you don't know it!

ex: "what is naive bayes"

Lin Programing

$$\left.\begin{aligned} &\tfrac{1}{2} x^T A x + b^T x \\ &\tfrac{1}{2} \|Ax - b\|^2 \\ &\sum_{i=1}^{N} z_i (w^T a_i - b)^2 \end{aligned}\right\} \text{Quadratic Functions}$$

$$\min_x \sum_{i=1}^{d} |x_i|$$

$\Downarrow$

Not Linear Constraint: (bad)

$v_i \geq \max \{ x_{ij} - x_i \}$

Linear Constraint: (good)

$v_i \geq x_i$

$v_i \geq -x_i$

non-smooth to smooth conversion

$\Downarrow$

$$\min_{x, v} \sum_{i=1}^{N} v_i \quad \text{subject to} \quad \underbrace{\begin{matrix} v_i \geq x_i \\ v_i \geq -x_i \end{matrix}}_{\text{linear constraints! good!}}$$

Norm equivalences: You can use them directly
(so maybe write them down!)

Hessian: $\nabla^2 f(x) = \sum_i a_i \overset{\text{outer product}}{a_i^T}$ "least squares Hessian" least squares: $f(x) = \tfrac{1}{2}\|Ax - b\|^2$

$d \times 1 \quad 1 \times d \quad = d \times d$

$$\begin{bmatrix} \sum_j a_{ij} a_{ij} & & \\ & & \\ & & \end{bmatrix}$$

$$\begin{bmatrix} a_1^T a_1 & a_1^T a_2 & \\ a_2^T a_1 & a_2^T a_2 & \\ & & \end{bmatrix} = A^T A$$

Lets say:

1×1 scalar

$$\nabla^2 f(x) = \sum_i a_i a_i^T d_i$$
$$= \sum_i (a_i d_i) a_i^T$$

$$\begin{bmatrix} \sum_j a_{1j} d_j a_{1j} & \sum_j a_{1j} d_j a_{2j} & \sum_j a_{1j} d_j a_{3j} \\ \sum_j a_{2j} d_j a_{1j} & \cdots & \\ \vdots & & \end{bmatrix}$$

$$\begin{bmatrix} a_{11} d_1 \\ a_{12} d_2 \\ a_{13} d_3 \\ \vdots \end{bmatrix} \qquad \begin{bmatrix} (a_1 \cdot d)^T a_1 & (a_1 \cdot d)^T a_2 \\ & \end{bmatrix}$$

$$= A^T \operatorname{diag}(d) A \qquad \begin{bmatrix} d_1 & & 0 \\ & d_2 & \\ 0 & & \ddots \end{bmatrix} \qquad \begin{bmatrix} d_1 (a_{11} \; a_{12} \cdots) \\ \\ \end{bmatrix}$$

a trick:

$$f(Ax)$$

$$\nabla f(Ax) = A^T \nabla f(x) \qquad \nabla f(x) = A^T \left[\frac{\partial f}{\partial x_i}\right]$$

$$\nabla^2 f(Ax) = A^T \underbrace{\operatorname{daig}(\nabla^2 f(x))}_{\sum_{i=1}^N f_i(a_i^T x)} A \qquad A^T \operatorname{daig}\left[\frac{\partial^2 f}{\partial x_i}\Big|_{Ax}\right] A$$

partial derivative evaluated at Ax

logistic

$$f(x) = \sum_{i=1}^N \log(1 + \exp(-b_i a_i^T x)) = f(Ax)$$

$$\nabla f(x) = A^T \left[y - \frac{y}{1 + \exp(b_i a_i^T x)} \right]$$

or something similar

$b\,\theta(w)$

$$\nabla^2 f(x) = A^T \operatorname{diag}(\text{something}) A$$

$\leftarrow \theta(w)(1 - \theta(w))$

write/memorize operations that preserve convexity!

- don't want to take hessians to proove convexity

another trick: (probably not on midterm, but you never know)

$$\tfrac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1$$

$$\Downarrow$$

$$\min_{x^+, x^-} \tfrac{1}{2}\|A(x^+ - x^-) - b^2 + \lambda \sum_j (x_j^+ + x_j^-) \qquad \text{quadratic}$$

$$\text{s.t.} \quad x_j^+ \geq 0 \qquad \text{linear}$$
$$x_j^- \geq 0$$

at solution

$$x = x^+ - x^-$$

$x \in \mathbb{R}^d$
$x^+ \in \mathbb{R}^d$
$x^- \in \mathbb{R}^d$

at solution

$$x^+ = x\, I(x > 0)$$
$$x^- = x\, I(x < 0)$$

derive co-ordinate descent

derive along a dimension (x_j^+, or x_j^-), set to zero, solve, ...

yadda yadda

linear constraints, same formulation as dual svm dual to primal

$f(Ax) + g(x)$

$-f^*(-y) - g(A^T y)$

if I now have optimal Y now? How to get X?

remember: $\sup_x \{ y^T x - f(Ax) \}$

x that solves this

no non-differentiable functions for the conj on midterm
(but maybe norms! or convert to smooth!)

$f(x) = \|x\|_p$

$f^*(y) = \begin{cases} \infty & \text{ie} \quad \|y\|_q \le 1 \end{cases}$

$f(x) = \frac{\lambda}{2} \|x\|_p^2 \Rightarrow f^*(y) = \frac{1}{2\lambda} \|y\|_q^2$

Fair game question:

$$\log(p(w|y,x)) : \underbrace{\sum_{i=1}^{N} \log(1 + \exp(-b_i a^T x))}_{\text{NLL}} + \underbrace{\frac{1}{2}\|w\|^2}_{\text{reg}}$$

possible question: show convex!

(log-likelihood)

$\exp(a+b)$
$= \exp(a)\exp(b)$

$\exp\left(-\sum_{i=1}^{N} \log(1+\exp(-b_i a^T x))\right)$

$\prod_i^N \exp(-\log(1+\exp(-b_i a^T x)))$

$\prod_i^N \frac{1}{1+\exp(-b_i a^T x)}$

(log-prior)

(regularizer)

$\exp(-\frac{\lambda}{2}\|w\|^2)$ ← gaussian

$w_j \sim N(0, \lambda^{-1})$

others:
L_1 laplacian

regularizer could be logistic!

logistic function

$\sigma(b_i a_i^T x)$ sigmoid

should be able to go both ways!

Converting to smooth, could ask:

$+ \lambda \|w\|_\infty$

$= \max_i \{ |w_i| \}$

$\max_i \{ \max \{ \cdots \} \}$

one variable that bounds
also bounds the absolute's

Multinomial could be on midterm!

$$\text{Mult}(x|\theta) = \theta_1^{I[x=1]} \theta_2^{I[x=2]} \theta_3^{I[x=3]}$$

$$\text{Dir}(\theta|a) \propto \theta_1^{a_1-1} \theta_2^{a_2-1} \theta_3^{a_3-1}$$

$$p(\theta|x,a) \propto p(x|\theta)p(\theta|a)$$

$$= \theta_1^{I[x=1]+a_1-1} \theta_2^{I[x=2]+a_2-1} \theta_3^{I[x=3]+a_3-1}$$

$$\theta_1^{(I[x=1]+a_1)-1} \theta_2^{(I[x=2]+a_2)-1} \theta_3^{(I[x=3]+a_3)-1}$$

$$p(\theta|x,a) \sim \text{Dir}(I[x=1]+a_1,\ I[x=2]+a_2,\ I[x=3]+a_3)$$ posterior, prior, same distribution family

$$p(x|n,\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \propto \theta^x(1-\theta)^{n-x}$$

$$p(\theta|a,b) = \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} \quad \propto \theta^{a-1}(1-\theta)^{b-1}$$

$$p(\theta|x,n,a,b) \propto \theta^x(1-\theta)^{n-x} \theta^{a-1}(1-\theta)^{b-1}$$

$$\Rightarrow \theta^{(x+a)-1}(1-\theta)^{(n-x+b)-1}$$

$$\theta|x,n,a,b \sim \text{Beta}(x+a,\ n-x+b)$$

Back to inf norm

$$\min_x \tfrac{1}{2}\|Ax-b\|^2 + \lambda\|x\|_\infty$$

$\Downarrow$ 1. write in terms of "max"

$$\min_x \tfrac{1}{2}\|Ax-b\|^2 + \lambda \max_j\{|x_j|\}$$

$$\min_x \tfrac{1}{2}\|Ax-b\|^2 + \lambda \max_j\{\max\{x_j, -x_j\}\}$$

$\Downarrow$ 2. upper bound max by linear variable

$$\min_{xv} \tfrac{1}{2}\|Ax-b\| + \lambda v$$ subject to $v \geq \max_j\{\max\{x_j, -x_j\}\}$

$\Downarrow$ 3. $v \geq \max\{a,b\} \Rightarrow v \geq a,\ v \geq b$

$$\min_{xv} \tfrac{1}{2}\|Ax-b\| + \lambda v$$ subject to

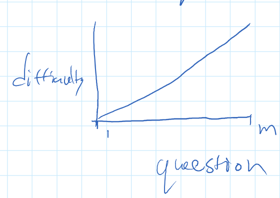
$v \geq x_j$

$v \geq -x_j$

Midterm Info all questions equally weighted, but harder questions later

likely:

8 questions



(bonus worth half a question)

Example Midterm Questions:

- show convex using convex properties (medium difficulty)
- bayes rule! (easy difficulty)
- converting non-smooth to smooth (medium difficulty)
- fenchel dual (medium-hard difficulty)