Quiz on Wednesday
Project proposal due Nov 3rd
HW3 due Nov 8th

TODAY
Logistic Regression
Gradient descent and Newton's Method
Separating hyperplanes (support vector machine)

In LR we assume $\eta(x) = P(Y=1 \mid X=x)$

$$= \frac{1}{1 + e^{-(w^T x + b)}}$$

Given data $(x_1, y_1), \ldots, (x_n, y_n)$ we estimate w and b using the method of Maximum Likelihood

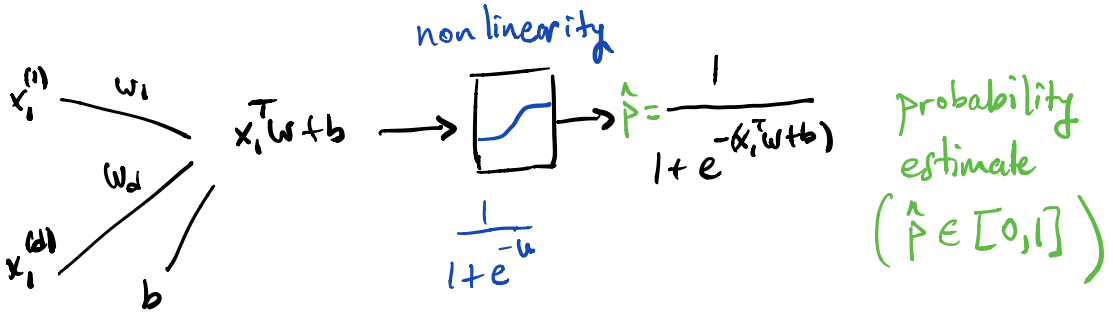
$$L(\theta) = \prod_{i=1}^{n} p(y_i \mid x_i, \theta)$$

$$= \prod_{i=1}^{n} p(x_i)^{y_i} \cdot (1 - p(x_i))^{1-y_i}$$

where $p(x_i) = P(Y=1 \mid X=x_i)$

$$-\log L(\theta) = -\sum_{i=1}^{n} y_i \cdot \log p(x_i) - (1-y_i) \cdot \log(1 - p(x_i))$$

(cross-entropy loss)



$$-\log L(\theta) = -\sum \log(1-p(x_i)) - \sum y_i \cdot \log \frac{p(x_i)}{1-p(x_i)}$$

$$= -\sum_{i=1}^{n} \log \frac{1}{1+e^{w^T x_i + b}} - \sum_{i=1}^{n} y_i (w^T x_i + b)$$

Define: $\tilde{x}_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix}$ $\theta = \begin{bmatrix} w \\ b \end{bmatrix}$ $\tilde{x}_i^T \theta = x_i^T w + b$

Exercise: If we define new labels $\tilde{y}_i \in \{-1, +1\}$, then

$$-\log L(\theta) = \sum_{i=1}^{n} \log\left(1 + e^{-\tilde{y}_i \theta^T \tilde{x}_i}\right) \qquad \left(\tilde{y}_i = 2 \cdot y_i - 1\right)$$

$y_i = 0 \Rightarrow \tilde{y}_i = -1$

$y_i = +1 \Rightarrow \tilde{y}_i = +1$

$-\log L(\theta)$: convex function.

$\nabla - \log L(\theta) = 0$ no closed form solution !

$$\nabla - \log L(\theta) = X^T \cdot (y - \hat{p}) \leftarrow \boxed{\text{HW 3}}$$

$$\nabla^2 - \log L(\theta) = X^T S X \qquad S \text{ is diagonal.}$$

Gradient Descent

$\min_{\theta} \ell(\theta)$

$$\theta_{t+1} = \theta_t - \alpha_t \cdot \nabla \ell(\theta_t)$$

α_t: step size

eg $\alpha = 0.1$

$$\ell(\theta_{t+1}) = \ell(\theta_t + \alpha \cdot v)$$

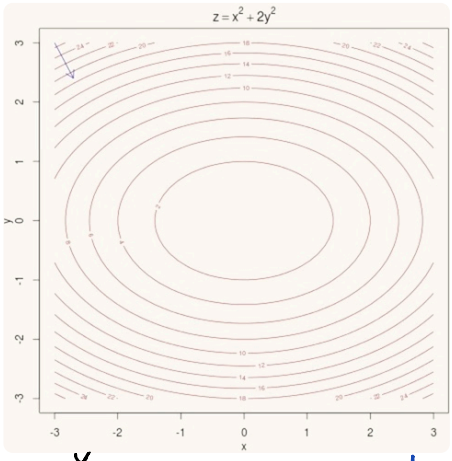
$$= \ell(\theta_t) + \alpha \cdot \langle \nabla \ell(\theta_t), v \rangle + O(\alpha^2)$$

$v^* = \underline{-\nabla \ell(\theta_t)}$ will minimize the above Taylor's approximation

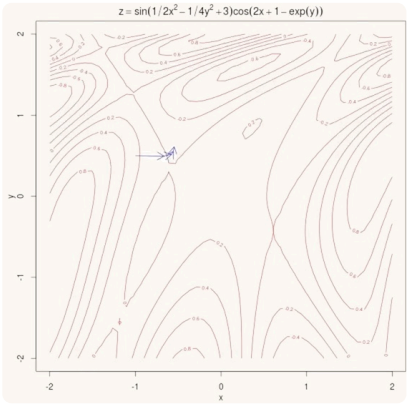
$\|\nabla \ell(\theta_t)\|_2$ 0 "

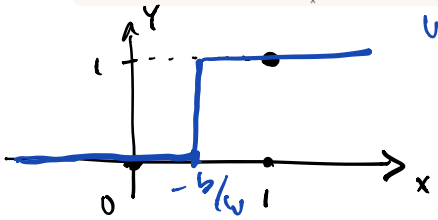
Cauchy-Schwartz: $\langle \nabla \ell(\theta_t), v \rangle \geq - \|\nabla \ell(\theta_t)\|_2 \cdot \|v\|_2$

Convex



Non-Convex





w, b

$\frac{1}{1+e^{wx+b}} \geq \frac{1}{2}$

$wx + b \geq 0$

$x \geq -b/w$

not unique !

$\frac{-\alpha \cdot b}{\alpha w} = -\frac{b}{w}$

linearly separable: $\hat{p} = y$. $\Rightarrow$ stationary point

$\nabla L(\theta) = 0.$

Newton's Method

$$\theta_{t+1} = \theta_t - \left(\nabla^2 \ell(\theta_t)\right)^{-1} \nabla \ell(\theta_t)$$

Second order Taylor's approximation.

$F(\theta)$

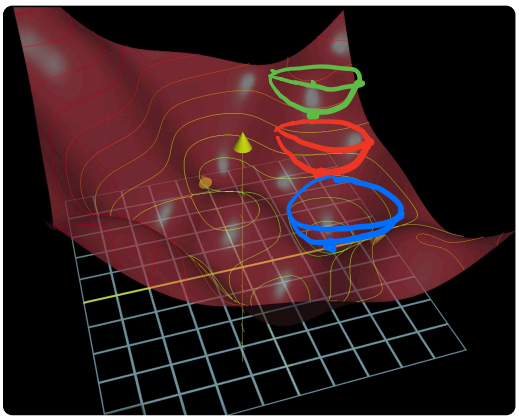
$$\ell(\theta) = \ell(\theta_t) + \langle \nabla \ell(\theta_t), \theta - \theta_t \rangle + \frac{1}{2} (\theta - \theta_t)^T \nabla^2 \ell(\theta_t)(\theta - \theta_t) + O(\|\theta - \theta_t\|_2^3)$$

$\nabla F(\theta) = 0$

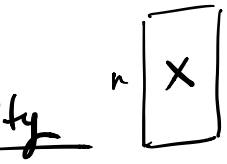
$$\nabla \ell(\theta_t) + \nabla^2 \ell(\theta_t)(\theta - \theta_t) = 0$$

$$\theta^+ = \theta_t - \left(\nabla^2 \ell(\theta_t)\right)^{-1} \cdot \nabla \ell(\theta_t)$$

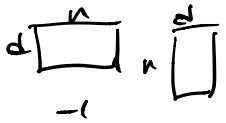
$= \theta_{t+1}$ (Newton's update)



	Complexity
GD	$O(nd)$



Newton $O(nd^2)$ Hessian $X^T S X$



$O(nd^2)$ $(\text{Hessian})^{-1}$ Gradient

direct methods

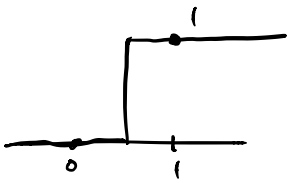
SVD

QR decomp.

Regularized Logistic Regression

Problem 1 : $n < d \Rightarrow (\text{Hessian})^{-1}$ doesn't exist!

Problem 2 :



linearly separable $\theta \to \infty$

no unique solution

Unique solutions

$$\min \quad L(\theta) + \lambda \|\theta\|_2^2 \qquad \text{L2 regularization}$$

$$\min \quad L(\theta) + \lambda \|\theta\|_1 \qquad \text{L1 regularization}$$

Exercise

LDA assumptions $\Rightarrow p(Y=1 \mid X=x) = \dfrac{1}{1+e^{-(w^T x + b)}}$

Separating hyperplanes (support vector machine)

LDA / QDA / NB $p(x)$ and $p(Y \mid X)$

Logistic $p(Y \mid X)$

SVM $w^T x + b \geq 0$ directly estimate $\theta = \begin{bmatrix} w \\ b \end{bmatrix}$

Vapnik's Principle:

> "When solving a problem of interest, do not solve a more general problem as an intermediate step."
>
> "Don't solve a harder problem than you have to"