CS 189 Final Note Sheet

Probabilistic Motivation for Least Squares

$y^{(i)} = \theta^\intercal x^{(i)} + \epsilon^{(i)} \quad \text{with} \quad \epsilon(i) \sim \mathcal{N}(0, \sigma^2)$

$\implies p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \theta^\intercal x^{(i)})^2}{2\sigma^2}\right)$

$\implies L(\theta) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y^{(i)} - \theta^\intercal x^{(i)})^2}{2\sigma^2}\right)$

$\implies l(\theta) = m \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^m (y^{(i)} - \theta^\intercal x^{(i)})^2$

$\implies \max_\theta l(\theta) \equiv \min_\theta \sum_{i=1}^m (y^{(i)} - \theta^\intercal x^{(i)})^2$

Gaussian noise in our data set $\{x^{(i)}, y^{(i)}\}_{i=1}^m$gives us least squares

$min_\theta ||X\theta - y||_2^2 \equiv \min_\theta \theta^\intercal X^\intercal X\theta - 2\theta^\intercal X^\intercal y + y^\intercal Y$

$\nabla_\theta l(\theta) = X^\intercal X\theta - X^\intercal y = 0 \implies \theta = (X^\intercal X)^{-1} X^\intercal y$

Gradient Descent: $\theta_{t+1} = \theta_t + \alpha(y_t^{(i)} - \theta^\intercal x_t^{(i)})x_t^{(i)}$

Least Squares Solution

$\min_x ||Ax - y||_2^2 \implies x^* = A^\dagger y$ min norm sol'n

Sol'n set: $x_0 + N(A) = x^* + N(A)$

$$A^\dagger = \begin{cases} (A^\intercal A)^{-1}A^\intercal & A \text{ full column rank} \\ A^\intercal(AA^\intercal)^{-1} & A \text{ full row rank} \\ V\Sigma^\dagger U^\intercal & \text{any } A \end{cases}$$

Logistic Regresion

Classify $y \in \{0, 1\} \implies$ model $p(y = 1|x) = h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

$\frac{dh_\theta}{d\theta} = (\frac{1}{1+e^{\theta^T x}})^2 e^{-\theta^T x} = \frac{1}{1+e^{\theta^T x}}\left(1 - \frac{1}{1+e^{-\theta^T x}}\right) = h_\theta(1 - h_\theta)$

$p(y|x;\theta) = (h_\theta(x))^y(1 - h_\theta(x))^{1-y} \implies$

$L(\theta) = \prod_{i=1}^m (h_\theta(x^{(i)}))^{y^{(i)}}(1 - h_\theta(x^{(i)}))^{1-y^{(i)}} \implies$

$l(\theta) = \sum_{i=1}^m y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)})) \implies$

$\nabla_\theta l = \sum_i (y^{(i)} - h_\theta(x^{(i)}))x^{(i)} = X^\intercal(y - h_\theta(X))$ (want max $l(\theta)$)

Stoch: $\boxed{\theta_{t+1} = \theta_t + \alpha(y_t^{(j)} - h_\theta(x_t^{(j)})x_t^{(j)}}$

Batch: $\boxed{\theta_{t+1} = \theta_t + \alpha X^\intercal(y - h_\theta(X))}$