Miles Benjamin
ID: 001776080

CS 6140 Midterm Cheat Sheet

Linear Regression

$J(\theta) = ||Y - X\theta||_2^2$

$\frac{dJ}{d\theta} = -2X^T(Y - X\theta)$

$\frac{d^2J}{d\theta} = 2X^TX$

Gradient Descent

While θ^k + 1 != θ^k:

θ^k+1 = θ^k - λ* dJ/dθ

Stochastic Gradient Descent:

Same as above but use mini batches of X and Y instead of the whole thing.

Huber Loss

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \le \delta, \\ \delta(|a| - \frac{1}{2}\delta), & \text{otherwise.} \end{cases}$$

Robust Regression

$J(\theta) = ||Y - X\theta||_2^2 - \lambda||\theta||_1$

Or $J(\theta) = ||Y - X\theta||_2^2 - \lambda||\theta||_2^2$

Uses Huber Loss as error function

Lasso Regression

$J(\theta) = ||Y - X\theta||_2^2 - \lambda||\theta||_1$

Ridge Regression

$J(\theta) = ||Y - X\theta||_2^2 - \lambda||\theta||_2^2$

K-fold Cross Validation

For i = 1 to k
  For λ∈Λ
    Compute θ using $\bigcup D^{n \neq i}$
    Compute error on D^i
    Pick λ with lowest error

Point Estimation

Maximum Likelihood Estimation

$P_\theta(x = 1) = \theta$ <- Hypothesis

$P_\theta(x) = \theta^x(1 - \theta)^{1-x}$ <- Bernoulli Dist

Probably Approximately Correct (PAC)

$P(|\theta_{mle} - \theta^0| \ge \varepsilon) \le 2e^{-N\varepsilon * \varepsilon}$

MAP Estimation

$P(\theta|D) = P(D|\theta)P(\theta)$

$\theta_{map} = (\Sigma x_i + \alpha - 1)/(N + \alpha + \beta - 2)$

For beta distribution prior

Generative Modeling for Classification: Naive Bayes

$1(y^i = 1) = 1 \; if \; y^i = 1, \; 0 \; else$

$aj = (1/N)\, \Sigma\, 1(y^i = j)$

$\theta_j^y = \Sigma\, 1(y^i = j) \,/\, N$

Discriminative Modeling for Classification: Logistic Regression

$J(\theta) = -\Sigma(y^i \, log(h_\theta(x^i)) + (1 - y^i)log(1 - h_\theta(x^i)))$

$P(y = 1|x) = h_\theta(x)$

Softmax Regression

$P(y(i) = k|x(i); \theta) = exp(\theta(k)\top x(i))\sum Kj = 1 exp(\theta(j)\top x(i))$

Perceptron Algorithm

1. Initialize w with random or zero
2. For t = 1 to T
   For i=1 to N
   $\overline{w} = \overline{w} - dJ/dt$

$dJ/dt = 0 \; if \; y^i\overline{w}^T x^i > 0$

$dJ/dt = -y^ix^i \; if \; y^i\overline{w}^T x^i < 0$

Functional and Geometric Margins

$x^i - z^i = w/||w||_2 \gamma^i$

$z^i = x^i - \gamma^i w/||w||_2$

$\gamma^i = (w^Tx^i + b)/(||w||_2)$

Support Vector Machines

Vanilla SVM (primal)

$min(w, b) \; ||w||_2 \quad s.t. \quad y^i(w^Tx^i + b) \ge 1$

Vanilla SVM (dual)

$max \; q(\alpha) = 1/2 \, \Sigma\Sigma\alpha_i\alpha_j y^i y^j < x^i, x^j > + \Sigma\alpha_i$

There will be a few points for which $\alpha_i > 0$, these are the points which define the line (the support vectors).

Max-Margin Classification

Lagrange Duality

$L(w, \alpha, \beta) = f(w) + \Sigma\alpha_i g_i(w) + \Sigma\beta_i h_i(w) \quad \alpha_i \ge 0, \; \beta_i \in R$

KKT Conditions

Does d* = p*?

Theorem: Assume {gi} are convex functions and {Hj} are affine functions. Also assume ∃θ s.t. { gi(θ) < 0}.

Then p* = d*

Bernoulli Distribution

X~ Bernoulli(p) 0<=p<=1

P(x=k) = {P if k = 1, 1-P if k= 0, 0 otherwise}

Binomial Distribution

X~B(n,p)

B(n, k ,p) = $\binom{n}{k} p^k (1-p)^{n-k}$

n -> # of trials, p -> 0 or 1, k -> # success

Gaussian Distribution

X~Normal(μ, σ^2)

Norm(x, μ, σ^2) = $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Laplace Distribution

X ~ Laplace(μ, b)

Lap(x, μ, b) = $\frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$