CS 760 Midterm Note Sheet

Overview

$\mathcal{X} \to$ input space
$\mathcal{Y} \to$ output space
$\mathcal{H} \to$ hypothesis class
Goal:
model $h \in \mathcal{H}$ that best approximates $f : \mathcal{X} \to \mathcal{Y}$

Empirical Risk Minimization

$\hat{f} = \arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(h(x^{(i)}), y^{(i)})$

k Nearest Neighbors

$\hat{y} = \sum_{i=1}^{k} y^{(i)}$ for k closest points
Distances:
Hamming: $d_H(x^{(i),x^{(j)}}) = \sum_{a=1}^{d} \mathbb{1}(x_a^{(i)} \neq x_a^{(j)})$
Euclidean: $d(x^{(i)}, x^{(j)}) = \left(\sum_{a=1}^{d}(x_a^{(i)} - x_a^{(j)})^2\right)^{-\frac{1}{2}}$
Manhattan: $d(x^{(i)}, x^{(j)}) = \sum_{a=1}^{d} |x_a^{(i)} - x_a^{(j)})^2|$

Information Theory

Entropy: $H(Y) = -\sum_{y\in Y} \mathbb{P}(y)\log_2(\mathbb{P}(y))$
Conditional: $H(Y|X) = \sum_{x\in X} \mathbb{P}(X = x)H(Y|X = x)$
Information Gain: $H(Y) - H(Y|X)$
Gain Ratio: $\frac{H(Y)-H(Y|S)}{H(S)}$

Evaluation

		Observed Class: True	Observed Class: False
Predicted Class	True	TP	FP
	False	FN	TN

Accuracy: $\frac{TP+TN}{TP+FP+TN+FN}$
Error: 1 - Accuracy
Precision: $\frac{TP}{TP+FP}$
Recall: $\frac{TP}{TP+FN}$
False Positive Rate: $\frac{FP}{TN+FP}$
ROC Curve: recall (TPR) vs. false positive rate
Precision/Recall Curve: downward sloping with asymptote

Decision Trees

Algorithm:
MakeSubtree:
Determine Candidate Splits
if stopping criteria: make leaf
else: *MakeSubtree* with best split
return subtree rooted at N

Linear Regression

Loss: $\ell(f_\theta) = \frac{1}{n}||\mathbf{X}\theta - y||_2^2$
Gradient: $\nabla_\theta = \frac{1}{n}(2\mathbf{X}^T\mathbf{X}\theta - 2\mathbf{X}^T y)$
Solution: $\theta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$
Ridge Loss: $\ell(f_\theta) = \frac{1}{n}||\mathbf{X}\theta - y||_2^2 + \lambda||\theta||_2^2$
Ridge Sol: $\theta = (\mathbf{X}^T\mathbf{X} + \lambda nI)^{-1}\mathbf{X}^T y$

Logistic Regression

Sigmoid: $\sigma(z) = \frac{1}{1+\exp(-z)} = \frac{\exp(z)}{1+\exp(z)} \in (0, 1)$
Properties: $1 - \sigma(z) = \sigma(-z)$, $\sigma^{'}(z) = \sigma(z)(1 - \sigma(z))$
Cross Entropy Loss:

Estimation

Maximum Likelihood:
$\hat{\theta}_{MLE} = \arg\max_\theta \mathcal{L}(\theta; Y, X)$
$\mathcal{L}(\theta; Y, X) = \prod_{i=1}^{n} \mathbb{P}_\theta(y_i|x_i)$
Maximum a posteriori Probability:
$\hat{\theta}_{MAP} = \arg\max_\theta \prod_{i=1}^{n} p(x^{(i)}|\theta)p(\theta)$

Gradient Descent

Convergence Criteria:
1. Convex: $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$
2. Differentiable
3. Lipschitz-continuous: $\nabla^2 f(x) \preccurlyeq LI$

recall: ($B - A$ is positive semidefinite if $A \preccurlyeq B$)
recall: if C is positive semidefinite then $x^T Cx \geq 0\ \forall x$

Probability Distributions

Binomial: $\binom{x}{n}\theta^x(1-\theta)^{n-x}$
Normal: $(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$
Beta: $\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$
Gamma: $\frac{r^s}{\Gamma(s)}x^{s-1}e^{-rx}$
Exponential: $\lambda e^{-\lambda x}$

Naive Bayes

Assumes conditional independence of features:

$$\begin{aligned} P(X_1, \ldots, X_k, Y) &\propto P(X_1, \ldots, X_k \mid Y)P(Y) \\ &= \left(\prod_{k=1}^{K} P(X_k \mid Y)\right)P(Y) \end{aligned}$$

Perceptrons

$$\hat{y}(x) = \begin{cases} 1, & w^T x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Algorithm:
for index i:
if $y^{(i)}w^T x^{(i)} < 1$ (i.e prediction is wrong):
then $w_{t+1} = w_t + y^{(i)}x^{(i)}$
else $w_{t+1} = w_t$
Mistake Bound: $(2 + D(S)^2)\gamma(S)^{-2}$
where $D(S)$ is the max diameter and $\gamma(S)$ is the largest margin we can have with dataset S

Neural Networks

For a single internal node:
Input: x, Weights: w, Bias: b, Activation Function: r
Output: $r(w^T x + b)$ which feeds into the next layer
Gradient Components: $\frac{\partial l}{\partial w} = (\hat{y} - y)x$, $\frac{\partial l}{\partial x} = (\hat{y} - y)w$
2-Layer: $\frac{\partial l}{\partial w_{11}} = \frac{\partial l}{\partial a_{11}}\frac{\partial a_{11}}{\partial w_{11}^{(1)}} = (\hat{y}-y)w_{11}^{(2)}a_{11}(1-a_{11})x_1$