CS 760 Midterm Note Sheet

Overview

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

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	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: $\theta = (\mathbf{X}^T\mathbf{X} + \lambda n I)^{-1}\mathbf{X}^T y$

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 $x^T C x \geq 0 \; \forall x$