CSCI 416/516 Final Study Guide

Name:

Student ID:

---

1 Decision Tree

	Cloudy	Not Cloudy
Raining	24/100	1/100
Not Raining	25/100	50/100

- **Problem 1: Joint Entropy.**
  Suppose $X = \{\text{Raining, Not raining}\}$, $Y = \{\text{Cloudy, Not cloudy}\}$. What is the joint entropy, $H(X, Y)$?
- **Problem 2: Specific Conditional Entropy.**
  Following the setup from the previous question, what is the entropy of cloudiness Y, given that it is raining ($X = \text{raining}$)?
- **Problem 3: General Conditional Entropy.**
  Following the setup from the previous question, what is the entropy of cloudiness Y, given the variable X?
- **Problem 4: Information Gain.**
  How is the Information Gain $IG(Y|X)$, given the entropy of Y, $H(Y)$, and the conditional entropy $H(Y|X)$?
- **Problem 5: Information Gain.**
  if X is completely uninformative about Y, what is the value of $IG(Y|X)$?
- **Problem 6: Information Gain.**
  if X is completely informative about Y, what is the value of $IG(Y|X)$?
- **Problem 7: Overfitting.**
  What is overfitting in the context of decision trees?
- **Problem 8: Tree Components.**
  Explain the concepts of nodes, branches, leaves, and root node in a decision tree.

- **Problem 9: IG and Tree.**
  How is the Information Gain used to build a decision tree?
- **Problem 10: Pros and Cons.**
  What are the advantages and disadvantages of using the decision tree?

2 Ensemble Learning

- **Problem 1: Weak Learners.**
  What are weak learners in the context of AdaBoost? Provide examples of common weak learners.
- **Problem 2: Misclassification.**
  Discuss the concept of misclassification rate and its role in AdaBoost. How are weights adjusted for misclassified samples?
- **Problem 3: Decision Stumps.**
  Discuss the relationship between AdaBoost and decision stumps. Why are decision stumps often chosen as weak learners in AdaBoost?
- **Problem 4: Sample weights.**
  What does the weight $w_{t,i}$ of a given sample x_i mean in the context of AdaBoost?
- **Problem 5: Objective function.**
  Explain what this objective function does, in the context of AdaBoost.

$$\mathcal{J}_{\text{reg}}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} w_i[y_1 \log h_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + (1-y_i)\log(1-h_{\boldsymbol{\theta}}(\boldsymbol{x}_i))] + \lambda||\boldsymbol{\theta}_{[1:d]}||_2^2 \tag{1}$$

3 Multilayer Perceptrons

- **Problem 1: Backpropagation.**
  Discuss the significance of the learning rate in the backpropagation algorithm. What can happen if it is set too high or too low?
- **Problem 2: Architecture.**
  Describe the typical architecture of a Multilayer Perceptron. Include details about the input layer, hidden layers, and the output layer.
- **Problem 3: Learning Rate.**
  Explain the significance of the learning rate in the training process of an MLP. What are the potential effects of setting it too high or too low?
- **Problem 4: Gradient Descent.**
  What's the relationship between gradient descent and backpropagation?
- **Problem 5: Activation Functions.**
  Explain the role of activation functions in MLPs.

4 Convolutional Neural Networks

- **Problem 1: Convolution.**
  What is the value given a vector a = [2, -1, 1] convolved by a filter b = [1, 1, 2]?

- **Problem 2: Activation Functions.**
  Given the answer from the previous question, what is the outcome after you apply ReLU on it?

- **Problem 3: Kernels.**
  What is the purpose of the filters/kernels used in a convolutional layer?

- **Problem 4: Overfitting.**
  What methods can be used to prevent overfitting in CNNs?

- **Problem 5: Pooling.**
  What is the purpose of pooling layers in a CNN? Compare and contrast Max Pooling and Average Pooling.

5 Attention & Transformers

- **Problem 1: Self-Attention.**
  Explain the concept of self-attention in transformers.

- **Problem 2: Multi-Head Attention.**
  What is multi-head attention in transformers?

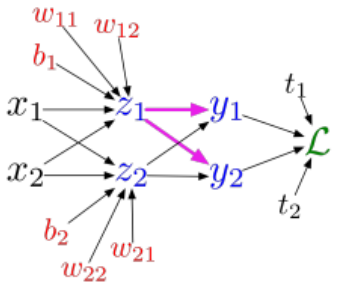
- **Problem 3: Encoders and Decoders.**
  Compare the roles of the encoder and decoder in a transformer model. How are they similar and different?

- **Problem 4: Attention.**
  Describe how the concept of attention, as used in transformers, can be applied to domains other than language processing, such as image or video analysis.

- **Problem 5: Application.**
  Suppose you want to use transformers for multiclass classification. How should you modify the existing transformer architecture to achieve this goal?

6 Miscallenous

Multclass logistic regression



$$z_\ell = \sum_j w_{\ell j} x_j + b_\ell$$

$$y_k = \frac{e^{z_k}}{\sum_\ell e^{z_\ell}}$$

$$\mathcal{L} = -\sum t_k \log y_k$$

- **Problem 1: Backpropagation.**
  What does the back pass look like, given the illustrated forward pass?