# CS540 Introduction to Artificial Intelligence

## Lecture 6

Young Wu

Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

July 6, 2022

# Hat Game

## Quiz

- 5 kids are wearing either green or red hats in a party: they can see every other kid's hat but not their own.
- Dad said to everyone: at least one of you is wearing green hat.
- Dad asked everyone: do you know the color of your hat?
- Everyone said no.
- Dad asked again: do you know the color of your hat?
- Everyone said no.
- Dad asked again: do you know the color of your hat?
- Some kids (at least one) said yes.
- No one lied. How many kids are wearing green hats?
- A: 1... B: 2... C: 3... D: 4... E: 5

# Hat Game Diagram

## Discussion

# Discussion Grades

## Admin

- The Discussion grades for week 1 is still computed incorrectly: will try to fix tonight.
- The past exam links are fixed, please refresh the pages.
- The list of relavent past exam questions are on the $Q1$, $Q2$, etc, pages.
- Please do NOT start the homework that are not announced yet! Especially please do not share solutions to those.

# Discussion Posts

## Admin

- Complete slides are listed as "Blank Slides".
- The slides used during the lectures are listed as "Blank Slides with Blank Pages for Quiz Questions".
- Shared solutions should explain how you get the solutions, a post will not get points if it just says "according to the hints, ...".
- Link to MyScript and Maple App on $W1$ page.

# Axes Aligned Decision Boundary

## Motivation

# Decision Tree

## Description

- Find the feature that is the most informative.
- Split the training set into subsets according to this feature.
- Repeat on the subsets until all the labels in the subset are the same.

# Binary Entropy

## Definition

- Entropy is the measure of uncertainty.
- The value of something uncertain is more informative than the value of something certain.
- For binary labels, $y_i \in \{0, 1\}$, suppose $p_0$ fraction of labels are 0 and $1 - p_0 = p_1$ fraction of the training set labels are 1, the entropy is:

$$\begin{aligned} H(Y) &= p_0 \log_2 \left(\frac{1}{p_0}\right) + p_1 \log_2 \left(\frac{1}{p_1}\right) \\ &= -p_0 \log_2 (p_0) - p_1 \log_2 (p_1) \end{aligned}$$

# Entropy

## Definition

- If there are $K$ classes and $p_y$ fraction of the training set labels are in class $y$, with $y \in \{1, 2, ..., K\}$, the entropy is:

$$H(Y) = \sum_{y=1}^{K} p_y \log_2 \left( \frac{1}{p_y} \right)$$

$$= - \sum_{y=1}^{K} p_y \log_2 (p_y)$$

# Entropy

## Quiz

- Running from You-Know-Who, Harry enters the CS building on the 1st floor. He flips a fair coin: if it is heads he hides in room 1325; otherwise, he climbs to the 2nd floor. In that case, he flips the coin again: if it is heads he hides in CSL; otherwise, he climbs to the 3rd floor and hides in 3331. What is the entropy of Harry's location?

# Entropy Math

## Quiz

# Entropy 2

## Quiz

- A bag contains a red ball, a green ball, a blue ball, and a black ball. Randomly draw a ball from the bag with equal probability. What is the entropy of the outcome?
- $A : 1$
- $B : \log_2(3)$
- $C : 1.5$
- $D : 2$
- $E : I$ don't understand entropy

# Conditional Entropy

## Definition

- Conditional entropy is the entropy of the conditional distribution. Let $K_X$ be the possible values of a feature $X$ and $K_Y$ be the possible labels $Y$. Define $p_x$ as the fraction of the instances that are $x$, and $p_{y|x}$ as the fraction of the labels that are $y$ among the ones with instance $x$.

$$H(Y|X=x) = -\sum_{y=1}^{K_Y} p_{y|x} \log_2 \left(p_{y|x}\right)$$

$$H(Y|X) = \sum_{x=1}^{K_X} p_x H(Y|X=x)$$

# Aside: Cross Entropy

## Definition

- Cross entropy measures the difference between two distributions.

$$H(Y, X) = -\sum_{z=1}^{K} p_{Y=z} \log_2 (p_{X=z})$$

- It is used in logistic regression to measure the difference between actual label $Y_i$ and the predicted label $A_i$ for instance $i$, and at the same time, to make the cost convex.

$$H(Y_i, A_i) = -y_i \log(a_i) - (1 - y_i) \log(1 - a_i)$$

# Information Gain

## Definition

- The information gain is defined as the difference between the entropy and the conditional entropy.

$$I(Y|X) = H(Y) - H(Y|X).$$

- The larger than information gain, the larger the reduction in uncertainty, and the better predictor the feature is.

# Splitting Discrete Features

## Definition

- The most informative feature is the one with the largest information gain.

$$\underset{j}{\operatorname{argmax}}\, I\left(Y|X_j\right)$$

- Splitting means dividing the training set into $K_{X_j}$ subsets.

$$\left\{(x_i, y_i) : x_{ij} = 1\right\}, \left\{(x_i, y_i) : x_{ij} = 2\right\}, ..., \left\{(x_i, y_i) : x_{ij} = K_{X_j}\right\}$$

# Splitting Continuous Variables Diagram

## Definition

# ID3 Algorithm (Iterative Dichotomiser 3)

## Description

- Find the feature that is the most informative.
- Split the training set into subsets according to this feature.
- Repeat on the subsets until all the labels in the subset are the same.

# Pruning Diagram

## Discussion

# Bagging Diagram

## Discussion

# Boosting Diagram

## Discussion

# $K$ Nearest Neighbor

## Description

- Given a new instance, find the $K$ instances in the training set that are the closest.
- Predict the label of the new instance by the majority of the labels of the $K$ instances.

# Distance Function

## Definition

- Many distance functions can be used in place of the Euclidean distance.

$$\rho\left(x, x'\right) = \left\|x - x'\right\|_2 = \sqrt{\sum_{j=1}^{m}\left(x_j - x'_j\right)^2}$$

- An example is Manhattan distance.

$$\rho\left(x, x'\right) = \sum_{j=1}^{m}\left|x_j - x'_j\right|$$

# Manhattan Distance Diagram

## Definition

# 1 Nearest Neighbor

## Quiz

- Find the 1 Nearest Neighbor label for $\begin{bmatrix} 3 \\ 6 \end{bmatrix}$ using Manhattan distance.

| | | | | | |
|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 3 | 5 | 2 |
| $x_2$ | 1 | 7 | 3 | 4 | 5 |
| $y$ | 0 | 1 | 1 | 0 | 0 |

- $A : 0$
- $B : 1$

# 3 Nearest Neighbor

## Quiz

- Find the 3 Nearest Neighbor label for $\begin{bmatrix} 3 \\ 3 \end{bmatrix}$ using Manhattan distance.

| | | | | | |
|---|---|---|---|---|---|
| $x_1$ | 1 | 1 | 3 | 5 | 2 |
| $x_2$ | 1 | 7 | 3 | 4 | 5 |
| $y$ | 0 | 1 | 1 | 0 | 0 |

- $A : 0$
- $B : 1$

# $K$ Fold Cross Validation

## Discussion

- Partition the training set into $K$ groups.
- Pick one group as the validation set.
- Train the model on the remaining training set.
- Repeat the process for each of the $K$ groups.
- Compare accuracy (or cost) for models with different hyperparameters and select the best one.

# 5 Fold Cross Validation Example

## Discussion

# Leave One Out Cross Validation

## Discussion

- If $K = n$, each time exactly one training instance is left out as the validation set. This special case is called Leave One Out Cross Validation (LOOCV).

# Cross Validation

## Quiz

- Given the following training data. What is the 2 fold cross-validation accuracy if 1 nearest neighbor classifier with Manhattan distance is used? The first fold is the first five data points.

| $x$ | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 1 | 2 | 3 | 3 | 2 | 2 | 3 | 3 | 2 | 1 |

# Cross Validation 2

## Quiz

- Given the following training data. What is the 10 fold cross-validation (LOOCV) accuracy if 1 nearest neighbor classifier with Manhattan distance is used?

| $x$ | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 1 | 2 | 3 | 3 | 2 | 2 | 3 | 3 | 2 | 1 |

- $A$ : 20 percent, B: 40, C: 60, D: 80, E: 100