CS540 Introduction to Artificial Intelligence
Lecture 6

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

June 29, 2021

Choose C

Admin

- A:
- B:
- C: Choose this.
- D:
- E:

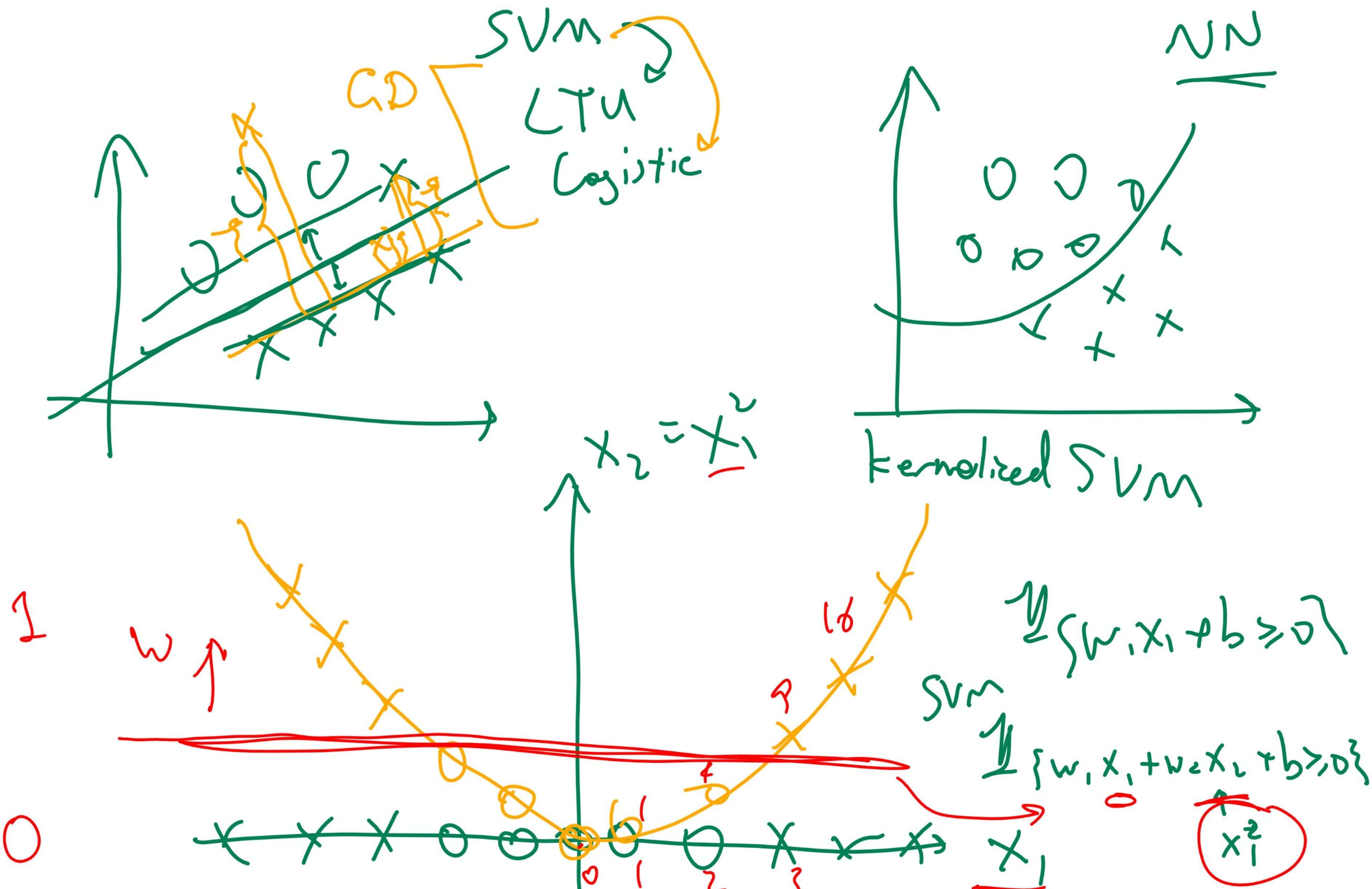
Remind Me to Start Recording

Admin

- The messages you send in chat will be recorded: you can change your Zoom name now before I start recording.

Kernel Trick 1*D* Diagram

Motivation



Kernelized SVM

$$\underbrace{b + w_1 x_1 + w_2 x_2 + w_3 x_3 = 0}_{\text{plane.}}$$

Definition

- With a feature map φ, the SVM can be trained on new data points $\{(\varphi(x_1), y_1), (\varphi(x_2), y_2), ..., (\varphi(x_n), y_n)\}$.
- The weights w correspond to the new features $\varphi(x_i)$.
- Therefore, test instances are transformed to have the same new features.

$$\hat{y}_i = \mathbb{1}_{\{w^T \varphi(x_i) \geqslant 0\}}$$

$$\varphi(x_i) \to \begin{pmatrix} x_i \\ x_i^2 \end{pmatrix}$$

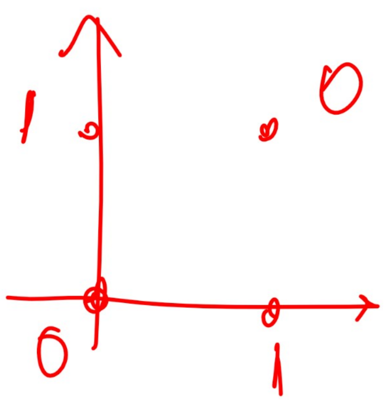
Kernel Trick for XOR

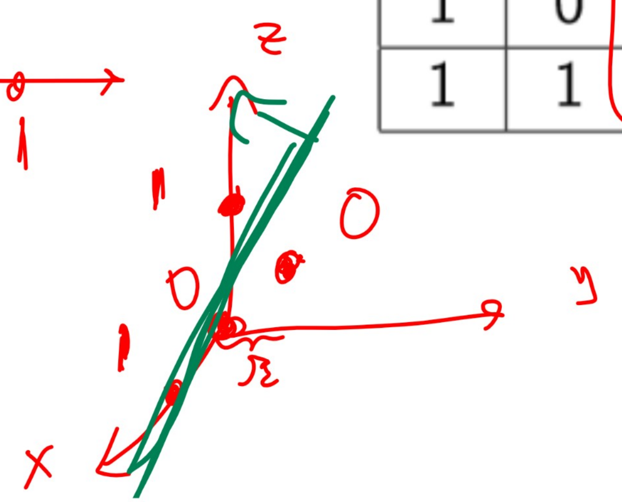
Quiz

- March 2018 Final Q17
- SVM with quadratic kernel $\varphi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ can correctly classify the following training set?

training

x_1	x_2	y	x_1^2	$\sqrt{2}x_1x_2$	x_2^2
0	0	0	0	0	0
0	1	1	0	0	1
1	0	1	1	0	0
1	1	0	1	$\sqrt{2}$	1





SVM plane

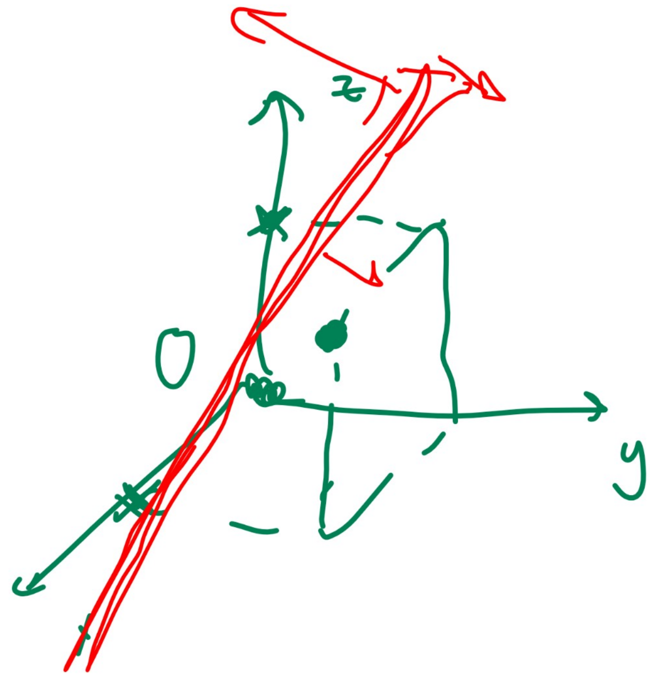
Kernel Trick for XOR

Quiz

Q2

- SVM with kernel $\varphi(x) = (x_1, x_1 x_2, x_2)$ can correctly classify the following training set?

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0



- A: True.
- B: False.

Kernel Matrix

Definition

not related
to features.

- The feature map is usually represented by a $n \times n$ matrix K called the Gram matrix (or kernel matrix).

$$K_{ii'} = \boxed{\varphi(x_i)^T \varphi(x_{i'})}$$

p.d.
sym.

Examples of Kernel Matrix

Definition

- For example, if $\varphi(x) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$, then the kernel matrix can be simplified.

$$K_{ii'} = \left(x_i^T x_{i'}\right)^2$$

- Another example is the quadratic kernel $K_{ii'} = \left(x_i^T x_{i'} + 1\right)^2$. It can be factored to have the following feature representations.

$$\varphi(x) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1\right)$$

Examples of Kernel Matrix Derivation

Definition

n

try $x_i = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ $\quad x_j' = \begin{pmatrix} x_1' \\ x_2' \end{pmatrix}$

x_{i_1}

x_{i_2}

$$K = (x^T x' + 1)^2$$

$$= \left((x_1, x_2)\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} + 1 \right)^2$$

$$= (x_1 x_1' + x_2 x_2' + 1)^2$$

$$= x_1^2 x_1'^2 + x_2^2 x_2'^2 + 1 \cdot 1 + (\sqrt{2} x_1 x_2)(\sqrt{2} x_1' x_2') + (\sqrt{2} x_1)(\sqrt{2} x_1') + (\sqrt{2} x_2)(\sqrt{2} x_2')$$

$$= \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \\ \sqrt{2} x_1 \end{pmatrix}^T \begin{pmatrix} x_1'^2 \\ x_2'^2 \\ \sqrt{2} x_1' x_2' \end{pmatrix} = \phi^T(x_i)\, \phi(x)$$

$(\sqrt{2}x_2$ $\begin{pmatrix} \sqrt{2}x_1 \\ \sqrt{2}x_2 \end{pmatrix}$

Popular Kernels

Discussion

- Other popular kernels include the following.

1. Linear kernel: $K_{ii'} = x_i^T x_{i'}$ ✓
2. Polynomial kernel: $K_{ii'} = \left(x_i^T x_{i'} + 1\right)^d$ ✓ $\phi^T \phi$
3. Radial Basis Function (Gaussian) kernel:

$$\boxed{K_{ii'} = \exp\left(-\frac{1}{\sigma^2}\left(x_i - x_{i'}\right)^T \left(x_i - x_{i'}\right)\right)}$$

$\phi^T \phi$

- Gaussian kernel has infinite-dimensional feature representations. There are dual optimization techniques to find w and b for these kernels.

in SS.

Kernel Matrix

Quiz

- Fall 2009 Final Q2
- What is the feature vector $\varphi(x)$ induced by the kernel $K_{ii'} = \exp(x_i + x_{i'}) + \sqrt{x_i x_{i'}} + 3$?
- A: $\left(\exp(x), \sqrt{x}, 3\right)$
- B: $\left(\exp(x), \sqrt{x}, \sqrt{3}\right)$
- C: $\left(\sqrt{\exp(x)}, \sqrt{x}, 3\right)$
- D: $\left(\sqrt{\exp(x)}, \sqrt{x}, \sqrt{3}\right)$
- E: None of the above

Q3

$\boxed{e^{x_i}}\ \boxed{e^{x_{i'}}} + \sqrt{x_i}\sqrt{x_{i'}} + \boxed{\sqrt{3}}\,\boxed{\sqrt{3}}$

$x_i \quad x_{i'}$

$K_{ii'} = \varphi(x_i)^T \varphi(x_{i'})$

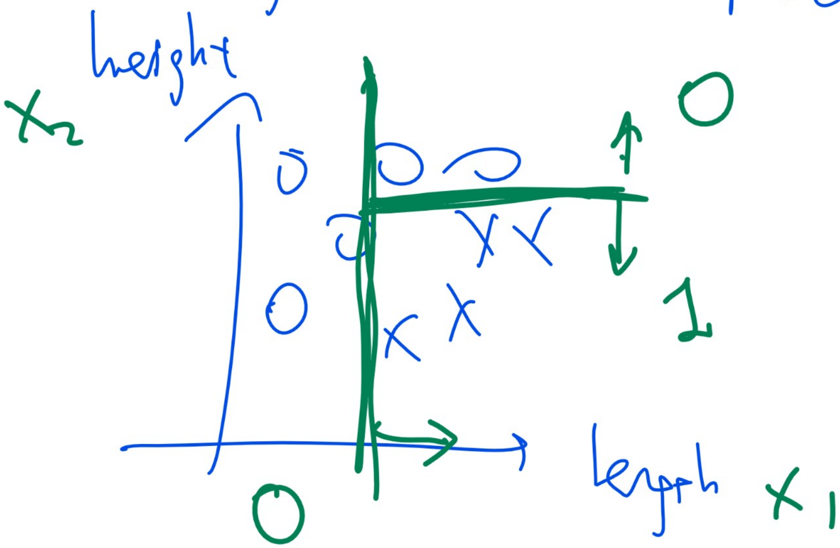
Kernel Matrix Math

Quiz

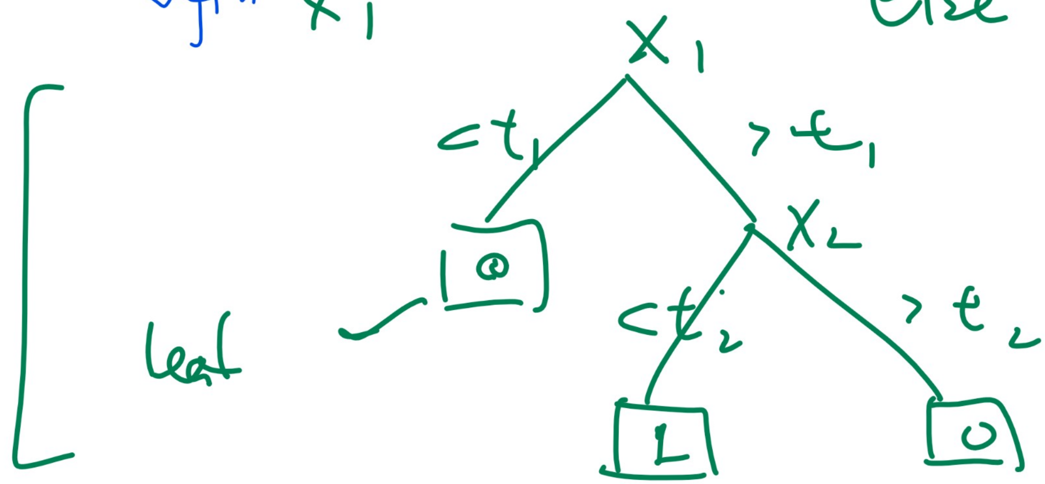
A Decision Tree

Motivation

weights cannot be explained

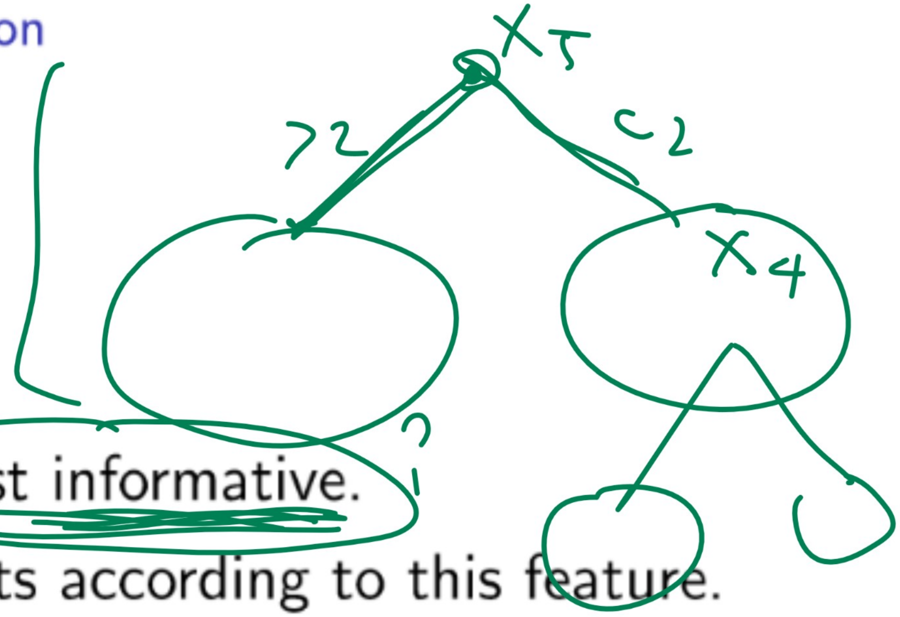


if $x_1 < t_1$
  return label 0
else
  if $x_2 > t_2$
    return 0
  else $x_2 \le t_2$
    return 1



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

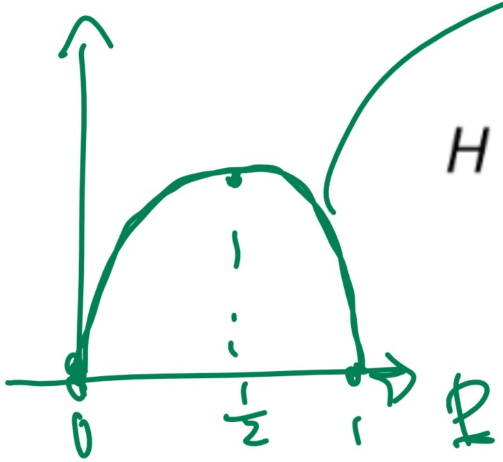
Definition

M4 Q9 on exam

- Entropy is the measure of uncertainty.
- The value of something uncertain is more informative than the value of something certain.

M3 Q9 on exam.

- 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:



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

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

- Fall 2010 Final Q10
- 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?
- A: 0.75
- B: 1
- C: 1.5
- D: 1.75
- E: None of the above.

1 2 3

$\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{4}$

$$-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{4}\log_2\frac{1}{4}$$

$$\frac{1}{2}+\frac{1}{2}+\frac{1}{2}$$

Entropy Math

Quiz

Entropy 2

Quiz

Q4

- 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 $= 4 \cdot \frac{1}{2}$
- E: 4

$$H(Y) = -\sum_{i=1}^{4} p_i \log_2 p_i$$

$\frac{1}{4}$ $\frac{1}{4}$

$p_1 \quad \frac{1}{4}$

$$\boxed{\log_2(a) = \frac{\log(a)}{\log(2)}}$$

$$\log_2 \frac{1}{4} = -\log_2 4 = -\log_2 2^2 = -2\log_2 2 = -2$$

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.

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

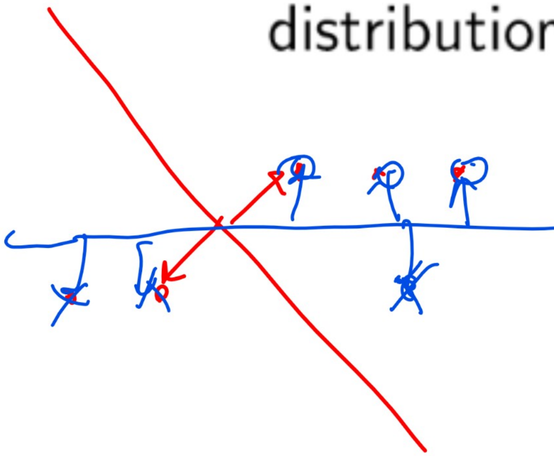
$$H\left(Y|X\right) = \sum_{x=1}^{K_X} p_x H\left(Y|X=x\right)$$

Aside: Cross Entropy

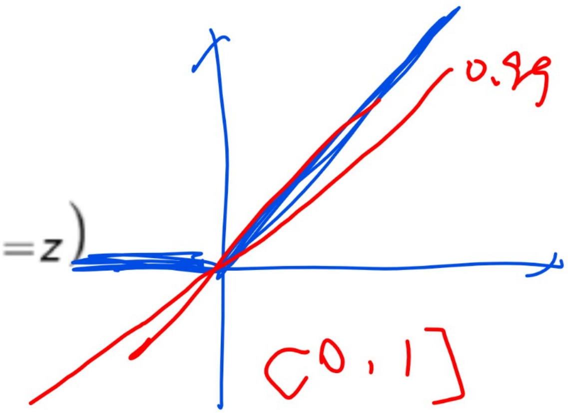
Definition

- Cross entropy measures the difference between two distributions.

max{x, 0}



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



- 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.

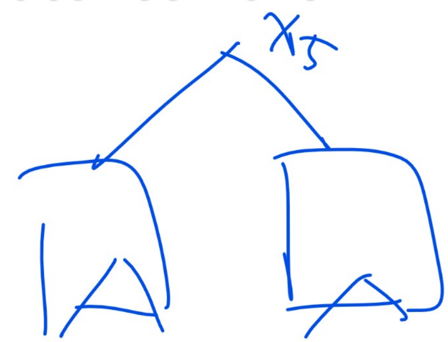
$$H(Y_i, A_i) = -y_i \log(a_i) - (1 - y_i) \log(1 - a_i)$$

Information Gain

Definition

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

$$\underset{X}{\arg\max}\; I(Y|X) = H(Y) - H(Y|X).$$



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

Information Gain Example

Quiz

- It has a house with many doors. A random door is about to be opened with equal probability. Doors 1 to 3 have monsters that eat people. Doors 4 to 6 are safe. With sufficient bribe, Pennywise will answer your question "Will door 1 be opened?" What's the information gain (also called mutual information) between Pennywise's answer and your encounter with a monster?

Information Gain Example

Quiz

- It has a house with many doors. A random door is about to be opened with equal probability. Doors 1 to 2 have monsters that eat people. Doors 3 to 4 are safe. With sufficient bribe, Pennywise will answer your question "Will door 1 be opened?". What's the information gain (also called mutual information) between Pennywise's answer and your encounter with a monster? Let $H_3 = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)$.
- A: $1 - \frac{1}{4} \cdot 0 - \frac{3}{4} \cdot 0$
- B: $1 - \frac{1}{4} \cdot 0 - \frac{3}{4} \cdot H_3$
- C: $1 - \frac{1}{4} \cdot H_3 - \frac{3}{4} \cdot 0$
- D: $1 - \frac{1}{4} \cdot H_3 - \frac{3}{4} \cdot H_3$

Splitting Discrete Features

Definition

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

$$\arg\max_j I\left(Y \middle| X_j\right)$$

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

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

Pruning Diagram

Discussion

Bagging and 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.

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

1 Nearest Neighbor

Quiz

- Spring 2018 Midterm Q7
- 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