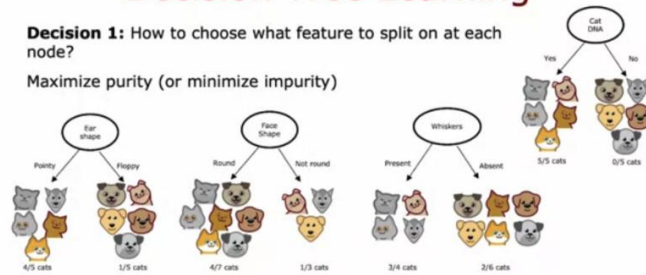


Decision Tree Learning

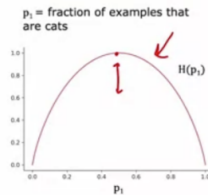
Decision 1: How to choose what feature to split on at each node?

Maximize purity (or minimize impurity)



2. Tome un aprendizaje de árbol de decisión para clasificar entre correo electrónico spam y no spam. Hay 20 ejemplos de entrenamiento en la nota raíz, que comprenden 10 correos electrónicos de spam y 10 de no spam. Si el algoritmo puede elegir entre cuatro características, que dan lugar a cuatro divisiones correspondientes, ¿cuál elegiría (es decir, cuál tiene mayor pureza)?
- ☐ División izquierda: 2 de 2 correos electrónicos son spam. División derecha: 8 de 18 correos electrónicos son spam.
 - ☐ División izquierda: 5 de 10 correos electrónicos son spam. División derecha: 5 de 10 correos electrónicos son spam.
 - ☒ División izquierda: 10 de 10 correos electrónicos son spam. División derecha: 0 de 10 correos electrónicos son spam.
 - ☐ División izquierda: 7 de 8 correos electrónicos son spam. División derecha: 3 de 12 correos electrónicos son spam.

Entropy as a measure of impurity



$$p_0 = 1 - p_1$$

$$H(p_1) = -p_1 \log_2(p_1) - p_0 \log_2(p_0) \\ = -p_1 \log_2(p_1) - (1 - p_1) \log_2(1 - p_1)$$

Note: "0 log(0)" = 0

1. Recordemos que la entropía se definió en clase como $H(p_1) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$, donde p_1 es la fracción de ejemplos positivos y p_0 la fracción de ejemplos negativos.

En un nodo dado de un árbol de decisión, 6 de 10 ejemplos son gatos y 4 de 10 no son gatos. ¿Qué expresión calcula la entropía $H(p_1)$ de este grupo de 10 animales?

- ☐ $-(0.6) \log_2(0.6) - (1 - 0.4) \log_2(1 - 0.4)$
- ☒ $-(0.6) \log_2(0.6) - (0.4) \log_2(0.4)$
- ☐ $(0.6) \log_2(0.6) + (1 - 0.4) \log_2(1 - 0.4)$
- ☐ $(0.6) \log_2(0.6) + (0.4) \log_2(0.4)$

Information gain

$$= H(p_1^{\text{root}}) - \left(w^{\text{left}} H(p_1^{\text{left}}) + w^{\text{right}} H(p_1^{\text{right}}) \right)$$

2. Recordemos que la información se definió de la siguiente manera:

$$H(p_1^{\text{root}}) - \left(w^{\text{left}} H(p_1^{\text{left}}) + w^{\text{right}} H(p_1^{\text{right}}) \right)$$

Antes de una división, la entropía de un grupo de 5 gatos y 5 no gatos es $H(5/10)$. Después de la división sobre una característica particular, un grupo de 7 animales (4 de los cuales son gatos) tiene una entropía de $H(4/7)$. El otro grupo de 3 animales (1 es un gato) y tiene una entropía de $H(1/3)$. ¿Cuál es la expresión para la ganancia de información?


☐ $H(0.5) - (7 * H(4/7) + 3 * H(1/3))$

☐ $H(0.5) - (H(4/7) + H(1/3))$

☒ $H(0.5) - \left(\frac{7}{10} H(4/7) + \frac{3}{10} H(1/3) \right)$

☐ $H(0.5) - \left(\frac{4}{7} * H(4/7) + \frac{4}{7} * H(1/3) \right)$

One hot encoding

Ear shape	Pointy ears	Floppy ears	Oval ears	Face shape	Whiskers	Cat
 Pointy	1	0	0	Round	Present	1
 Oval	0	0	1	Not round	Present	1
 Oval	0	0	1	Round	Absent	0
 Pointy	1	0	0	Not round	Present	0
 Oval	0	0	1	Round	Present	1
 Pointy	1	0	0	Round	Absent	1
 Floppy	0	1	0	Not round	Absent	0
 Oval	0	0	1	Round	Absent	1
 Floppy	0	1	0	Round	Absent	0
 Floppy	0	1	0	Round	Absent	0

3. Para representar 3 valores posibles para la forma de la oreja, puede definir 3 características para la forma de la oreja: orejas puntiagudas, orejas caídas, orejas ovaladas. Para un animal cuyas orejas no son puntiagudas, ni caídas, sino ovaladas, ¿cómo puede representar esta información como un vector de características?

☒ $[0, 0, 1]$

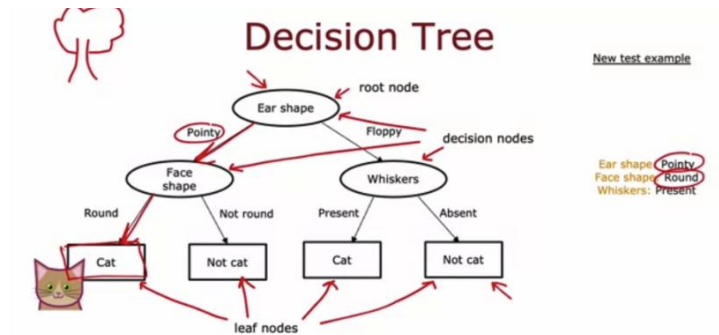
☐ $[1, 1, 0]$

☐ $[1, 0, 0]$

☒ $[0, 1, 0]$

5. ¿Cuáles de estos son los criterios utilizados habitualmente para decidir dejar de dividir? (Elija dos.)

- ☒ Cuando el árbol haya alcanzado una profundidad máxima
- ☐ Cuando la ganancia de información de las divisiones adicionales es demasiado grande
- ☒ Cuando el número de ejemplos en un nodo está por debajo de un umbral
- ☐ Cuando un nodo es 50% de una clase y 50% de otra clase (el valor más alto posible de entropía)



1. Basándose en el árbol de decisión mostrado en la conferencia, si un animal tiene orejas caídas, una forma de cara redonda y tiene bigotes, ¿predice el modelo que es un gato o que no es un gato?

☐ No es un gato

☒ cat