Home Work #2 - Ophere & Roni

Part 1: Theoretical Exercises (16 points)

1. Gini Impurity

In class, we defined the Gini impurity as

$$\varphi_{Gini}(p) = 1 - \sum_{j=1}^{k} p_j^2, \qquad p \in [0,1]^k,$$

where $p = (p_1, \ldots, p_k)$ represents class proportions in a set of instances. This means that $\sum_{j=1}^{k} p_j = 1$.

1. Prove that

$$\varphi_{Gini}(p) \leq 1 - 1/k.$$

Hint:

- Express the function $f : \mathbb{R}^{k-1} \to \mathbb{R}$:

$$f(p_1, \ldots, p_{k-1}) = \varphi_{Gini}(p_1, \ldots, 1 - \sum_{j=1}^{k}).$$

- Argue that f is bounded from above, hence it has a maximal value in $\mathbb{R}^{k-1}$.
- Solve the equation $\nabla f = 0$ and argue that the solution is unique.

(you do not have to follow the hint; all correct and clearly written solutions are acceptable)

Let Y_1 and Y_2 be two independent random variables, each represnting the class label of a randomly sampled instance from the set. Namely:

$$\Pr[Y_i = j] = p_j, \qquad i \in \{1, 2\}, \qquad j \in \{1 \ldots k\}.$$

2. Prove that Gini impurity is the probability that two randomly sampled instances (with replacement) from the set of instances have different class labels. Namley, that

$$\varphi_{Gini}(p) = \Pr[Y_1 \neq Y_2].$$

Question #1

1. $\varphi Gini(p) = 1 - \sum_{j=1}^{K} p_j^2 \qquad p \in [0,1]^K \qquad \left(\sum_{j=1}^{k} p_j = 1 \text{ sum of proportions}\right)$

Prove: $\varphi Gini(p) \leq 1 - \frac{1}{K}$

$\varphi Gini(p_1 \ldots p_K) = 1 - \sum_{j=1}^{K} p_j^2$ & $\sum_{j=1}^{K} p_j = 1$ → therefore $\sum_{j=1}^{K-1} p_j + p_K = 1$

$$p_K = 1 - \sum_{j=1}^{K-1} p_j$$

We'll define a function: $f: \mathbb{R}^{K-1} \to \mathbb{R}$ s.t. $f(p_1 \ldots p_K) = \varphi Gini(p_1 \ldots 1 - \sum_{j=1}^{K-1} p_j)$ (Hint)

We'll also define $S = \sum_{j=1}^{K-1} p_j$. Note that $p_1 + \ldots + p_{K-1} = \sum_{j=1}^{K-1} p_j \rightarrow 1 - \sum_{j=1}^{K-1} p_j + \sum_{j=1}^{K-1} p_j = 1$ ~ sum of proportions is 1, therefore f is well defined

· We will now prove f is bounded from above:

$$f(p_1 \ldots p_{K-1}) = \varphi Gini(\underbrace{p_1 \ldots 1 - \sum_{j=1}^{K-1} p_j}_{K-1 \text{ elements}}) = \varphi Gini(p_1 \ldots 1 - S) = 1 - \Big(\overbrace{\sum_{j=1}^{K-1} p_j^2 + \big(1 - \sum_{j=1}^{K-1} p_j\big)^2}^{1 \geq \text{sum of prop}^2 \geq 0}\Big) \leq 1 \rightarrow f \text{ is bounded from above by } 1.$$

By the definition of Gini

Note: $1-\left(\sum_{j=1}^{K-1} p_j^2+\left(1-\sum_{j=1}^{K-1} p_j\right)^2\right)$

Now we'll compute $\nabla f=0$ (gradient) to find the edge point:

We need to compute each partial derivative $\frac{\partial f}{\partial p}=0 \quad \forall 1\le i\le k$

$$\frac{\partial f}{\partial p_i}\, 1-\left(\sum_{j=1}^{K-1} p_j^2+\left(1-\sum_{j=1}^{K-1} p_j\right)^2\right)=0-\left(2p_i+2\left(1-\sum_{i=1}^{K-1} p_i\right)\cdot(-1)\right)=-2p_i+2-2\sum_{i=1}^{K-1} p_i \xrightarrow[=0]{\text{Compare to zero}} p_i+\sum_{i=1}^{K-1} p_i=1$$

$$\downarrow$$

$$\sum_{i=1}^{K} p_i=1$$

$$\downarrow$$

$$k\cdot p_i=1 \rightarrow p_i=\frac{1}{k} \quad \forall 1\le i\le k$$

Because $p_i=\frac{1}{k}$ is a unique solution there exists one edge point.

Using inequality Kaushy Shwartz we'll show that the edge point is a max point.

$$\varphi Gini(p_1 \dots p_K)=1-\sum_{j=1}^{K} p_j^2$$

Note: $\sum_{j=1}^{K} p_j=1$ & $p_j\ge 0$. In order to maximize Gini we need to minimize $\sum_{j=1}^{K} p_j^2$, under the constraint ~~tat~~ $\sum_{j=1}^{K} p_j$.

Reminder: For any two vectors $\rightarrow$ $(u\cdot v)^2\le |u|^2\cdot|v|^2 \qquad u,v\in\mathbb{R}^n$

We choose $u=\underbrace{(p_1 \dots p_K)}_{K \text{ times}}$ & $v=\underbrace{(1,1 \dots 1)}_{K \text{ times}}$

Kaushy shwartz inequality

$$\left(\sum_{j=1}^{K} p_j\times 1\right)^2\le\left(\sum_{j=1}^{K} p_j^2\right)\cdot\overbrace{\left(\sum_{j=1}^{K} 1^2\right)}^{K}$$

$$\left(\sum_{j=1}^{K} p_j\right)^2\le K\cdot\sum_{j=1}^{K} p_j^2$$

Sum $p_1 \dots p_K=1$

$$1\le K\cdot\sum_{j=1}^{K} p_j^2$$

$$\frac{1}{K}\le\sum_{j=1}^{K} p_j^2$$

*

$$\varphi Gini(p)=1-\sum_{j=1}^{K} p_j^2$$

$$\sum_{j=1}^{K} p_j^2=1-\varphi Gini(p)$$

$$\frac{1}{K}\le 1-\varphi Gini(p)$$

$$\varphi Gini(p)\le 1-\overbrace{\frac{1}{K}}^{=p_i}$$:)

Gini is less than or equal $1-p_i$ for $p_i=\frac{1}{K}$

2. Let Y_1 & Y_2 be two independent random variables. Each represent the class label of a randomly sampled instance from the set: $\Pr(Y_i = j) = p_j \quad i \in \{1,2\} \quad j \in \{1 \dots k\}$

Prove: gini impurity

$$\varphi_{Gini}(p) = \Pr[Y_1 \neq Y_2]$$

probability the two samples have different class labels.

$$\Pr[Y_1 \neq Y_2] = 1 - \Pr[Y_1 = Y_2] \text{ ([illegible])}$$

$$\Pr[Y_1 = j \wedge Y_2 = j] \overset{\text{independent}}{=} \Pr[Y_1 = j] \cdot \Pr[Y_2 = j] = p_j \cdot p_j = p_j^2$$

both samples have same label

$j \in \{1 \dots k\}$ so we'll compute probability for each of the possibilities:

$$\underbrace{\Pr \biguplus_{j=1}^{k} (Y_1 = j \wedge Y_2 = j) = \sum_{j=1}^{k} p_j^2}$$

$$\Pr[Y_1 \neq Y_2] = 1 - \Pr[Y_1 = Y_2] = 1 - \sum_{j=1}^{k} p_j^2 \overset{\text{definition}}{=} \varphi_{Gini}(p)$$

Question #2

2. Information Gain

In class we claimed that **information gain is always non-negative**. Here, we will prove this for the specific case of binary classification, where we have only two class labels.

Recall that information gain is defined as follows:

$$IG(S, A) \;=\; H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v)\,,$$

where S is a set of data instances, A is an attribute (ferature) with a finite set of possible values $Values(A)$, and H is the entropy function applied to the probability vector associated with the class frequencies. Assuming that there are only two class lables, the entropy can be expressed as follows:

$$H(S) \;=\; h(p_1) \;=\; -p_1 \log(p_1) - (1 - p_1)\log(1 - p_1)\,,$$

where p_1 is the frequency of the first label (and $1 - p_1$ is the frequency of the second label). Here, we adhere to the convention that $0 \cdot \log(0) = 0$ (as $\log(0)$ is undefined).

We start by examining the function $h()$, which is also called the *binary entropy function* (see plot below). One feature of this function is that it is concave. Concave functions satisfy the following property: for every $x_1, x_2 \in [0, 1]$ and for every $\lambda_1, \lambda_2 \in [0, 1]$ such that $\lambda_1 + \lambda_2 = 1$, we have:

$$h(\lambda_1 x_1 + \lambda_2 x_2) \geq \lambda_1 h(x_1) + \lambda_2 h(x_2). \tag{1}$$

1. Use the inequality in (1) to prove (by induction) a more general claim: for any $t \geq 2$ points $x_1 \dots x_t \in [0, 1]$, and t weights $\lambda_1 \dots \lambda_t \in [0, 1]$ such that $\sum_{j=1}^{t} \lambda_j = 1$, we have

$$h\left(\sum_{j=1}^{t} \lambda_j x_j\right) \geq \sum_{j=1}^{t} \lambda_j h(x_j)\,.$$

This inequality, which applies to all concave functions, is also called *Jensen's inequality*.

2. Use the inequality you proved above to prove that information gain is always non-negative (when there are only two classes).

1. Using $h(\lambda_1 x_1 + \lambda_2 x_2) \overset{*}{\geq} \lambda_1 h(x_1) + \lambda_2 h(x_2)$ $\quad \forall\ x_1, x_2 \in [0,1],\ \lambda_1, \lambda_2 \in [0,1]$ s.t. $\lambda_1 + \lambda_2 = 1$

we'll prove the claim by induction:

Base: For $t=2 \rightarrow$ $h\left(\sum_{j=1}^{2} \lambda_j x_j\right) = h\left(\lambda_1 x_1 + \lambda_2 x_2\right) \overset{*}{\geq} \lambda_1 h(x_1) + \lambda_2 h(x_2) = \sum_{j=1}^{2} \lambda_j h(x_j)$

Hypothesis: We assume that $h\left(\sum_{j=1}^{t-1} \lambda_j x_j\right) \geq \sum_{j=1}^{t-1} \lambda_j h(x_j)$

Step: We'll prove $h\left(\sum_{j=1}^{t} \lambda_j x_j\right) \geq \sum_{j=1}^{t} \lambda_j x_j \quad \forall\ t \geq 2$

Base case (2 points) and the induction hypothesis

$$h\left(\sum_{j=1}^{t} \lambda_j x_j\right) = h\left(\lambda_t x_t + \sum_{j=1}^{t-1} \lambda_j x_j\right) \geq \lambda_t h(x_t) + \sum_{j=1}^{t-1} \lambda_j h(x_j) = \sum_{j=1}^{t} \lambda_j h(x_j)$$

2. Using the inequality we'll prove the claim: $IG(S,A) \geq 0$ for 2 labels

data, finite set of attributes, need to prove

$$IG(S,A) = H(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} H(S_v) \geq 0$$

Given: $H(S) = h(p_1) = -\sum_{i=1}^{2} p_i \log(p_i) = -p_1 \log(p_1) - (1-p_1)\log(1-p_1)$

$$H(S) = h\left(\sum_{j=1}^{t} \lambda_j S_j\right) \geq \sum_{j=1}^{t} \lambda_j h(S_j) = \sum_{j=1}^{t} \frac{|S_j|}{|S|} H(S_j)$$

$1 \leq j \leq t$ values

$\lambda_j = \frac{|S_j|}{|S|}$

based on the jensen inequality

$$H(S) \geq \sum_{j=1}^{t} \frac{|S_j|}{|S|} H(S_j)$$

$$IG(S,A) = H(S) - \sum_{j=1}^{t} \frac{|S_j|}{|S|} H(S_j) \geq 0$$