Large Scale Machine Learning and Deep Learning
Review Questions 3

1. Which of the following is/are true about Bagging Trees and Boosting Trees?

   (a) In Bagging Trees, individual trees (weak learners) are independent of each other.
   (b) Bagging is a method for improving the performance by aggregating the results of weak learners.
   (c) In Boosting Trees, individual trees (weak learners) are independent of each other.
   (d) Boosting is a method for improving the performance by aggregating the results of weak learners.

   Answer: a, b, and d

---

2. Which of the following is/are true about individual tree in Random Forest?

   (a) Individual tree is built on a subset of the features.
   (b) Individual tree is built on all the features.
   (c) Individual tree is built on a subset of instances.
   (d) Individual tree is built on full set of instances.

   Answer: a and c

---

3. Ensemble model estimators (such as Random Forest) in Spark have a parameter called `featureSubsetStrategy`. What does it do?

   Answer: it determines the number of features to consider for splits at each node. Supported values are `auto`, `all`, `sqrt`, `log2`, `onethird`.

---

4. Explain why the entropy becomes zero when all class partitions are pure?

   Answer: in a Decision Tree, the entropy is defined by:

$$\texttt{entropy(D)} = -\sum_{\texttt{i}=\texttt{1}}^{\texttt{m}} \texttt{p}_\texttt{i} \log(\texttt{p}_\texttt{i})$$

   where $\texttt{p}_\texttt{i}$ is the probability that an instance in $\texttt{D}$ belongs to a class $\texttt{i}$, with $\texttt{m}$ distinct classes. If a partitions $\texttt{k}$ is pure, then $\texttt{p}_\texttt{k} = \texttt{1}$ (and thus $\texttt{p}_\texttt{i} = \texttt{0}$, for all $\texttt{i} \neq \texttt{k}$), and therefor we have $\texttt{entropy(D)} = -\texttt{1} \times \log(\texttt{1}) = \texttt{0}$.

---

5. Explain why the Gini impurity becomes zero when all class partitions are pure?

Answer: in a Decision Tree, the Gini impurity is defined by:

$$\mathtt{Gini}(\mathtt{D}) = 1 - \sum_{\mathtt{i}=1}^{\mathtt{m}} \mathtt{p}_{\mathtt{i}}^2$$

where $\mathtt{p}_{\mathtt{i}}$ is the probability that an instance in $\mathtt{D}$ belongs to a class $\mathtt{i}$, with $\mathtt{m}$ distinct classes. But, how the above formula measures an impurity? Imagine an experience with $\mathtt{m}$ possible output categories, in which category $\mathtt{i}$ has a probability of occurrence $\mathtt{p}_{\mathtt{i}}$ (where $\mathtt{i} = 1, \cdots \mathtt{m}$). Then, reproduce this experience two times and make these observations:

- the probability of obtaining two identical outputs of category $\mathtt{i}$ is $\mathtt{p}_{\mathtt{i}}^2$.
- the probability of obtaining two identical outputs, independently of their category, is $\sum_{\mathtt{i}=1}^{\mathtt{m}} \mathtt{p}_{\mathtt{i}}^2$.
- the probability of obtaining two different outputs is thus $1 - \sum_{\mathtt{i}=1}^{\mathtt{m}} \mathtt{p}_{\mathtt{i}}^2$.

The Gini impurity is simply the probability of obtaining two different outputs, which is an *impurity measure*. In the other direction, if we have a category $\mathtt{k}$ such that $\mathtt{p}_{\mathtt{k}} = 1$ (and thus $\mathtt{p}_{\mathtt{i}} = 0$, for all $\mathtt{i} \neq \mathtt{k}$) we have a Gini impurity $\mathtt{gini}(\mathtt{D}) = 1 - 1 = 0$, and we will always get two identical outputs (of category $\mathtt{k}$), which is a *pure* situation.