Feature preprocessing and generation with respect to models

Quiz, 4 questions

1 point

1.
Suppose we have a feature with all the values between 0 and 1 except few outliers larger than 1. What can help us to decrease outliers' influence on non-tree models?

- [ ] MinMaxScaler
- [x] Apply **np.sqrt(x)** transform to the data
- [x] Winsorization
- [x] Apply rank transform to the features
- [x] Apply **np.log1p(x)** transform to the data
- [ ] StandardScaler

2 points

2.
Suppose we fit a tree-based model. In which cases label encoding can be better to use than one-hot encoding?

- [x] When we can come up with label encoder, that assigns close labels to similar (in terms of target) categories
- [x] When categorical feature is ordinal
- [x] When the number of categorical features in the dataset is huge

1 point

3.

Suppose we fit a tree-based model on several categorical features. In which cases applying one-hot encoding can be better to use than label encoding?

- [x] If target dependence on the label encoded feature is very non-linear, i.e. values that are close to each other in the label encode feature correspond to target values that aren't close.
- [ ] When the feature have only two unique values

---

1 point

4.
Suppose we have a categorical feature and a *linear* model. We need to somehow encode this feature. Which of the following statements are true?

- ( ) Label encoding is always better than one-hot encoding
- ( ) One-hot encoding is always better than label encoding
- (•) Depending on the dataset either of label encoder or one-hot encoder could be better

---

- [ ] I, **Jiadai Zhao**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.

Learn more about Coursera's Honor Code

Submit Quiz