Introduction to ML CS M146 UCLA Spring Quarter 2022
Prof. Aditya Grover Thursday, 21th April 2022

Homework Set #2 : Decision trees, K-Nearest Neighbors and Kernel methods

Due 4th May 2022, Wednesday, before 11:59 pm

Submission instructions

- Submit your solutions electronically on the course Gradescope site as PDF files.
- If you plan to typeset your solutions, please use the LaTeX solution template. If you must submit scanned handwritten solutions, please use a black pen on blank white paper and a high-quality scanner app.
- You are required to paste the snapshots of your codes with the title 'Source Codes' at the end of your solutions pdf.

Problem 1 (Decision Trees)

You are stranded on a deserted island. Mushrooms of various types grow wildly all over the island, but no other food is anywhere to be found. Some of the mushrooms have been determined as poisonous and others as not (determined by your former companions' trial and error). You are the only one remaining on the island. You have the following data to consider.

Example	IsHeavy	IsSmelly	IsSpotted	IsSmooth	IsPoisonous
A	0	0	0	0	0
B	0	0	1	0	0
C	1	1	0	1	0
D	1	0	0	1	1
E	0	1	1	0	1
F	0	0	1	1	1
G	0	0	0	1	1
H	1	1	0	0	1
U	1	1	1	1	?
V	0	1	0	1	?
W	1	1	0	0	?

You know whether or not mushrooms A through H are poisonous, but you do not know about U through W. For the question a) - d), consider only mushrooms A through H.

(a) What is the entropy of IsPoisonous? **[1pt]**

(b) Which attribute should you choose as the root of a decision tree? **[1pt]**
Hints: You can figure this out by looking at the data without explicitly computing the information gain of all four attributes.

(c) What is the information gain of the attribute you chose in the previous question? **[5pts]**

(d) Using ID3 algorithm, build a decision tree to classify mushrooms as poisonous or not. **[5pts]**

(e) Classify mushrooms U, V and W using this decision tree as poisonous or not. **[3pts]**

Problem 2 (Entropy and Information)

The entropy of a Bernoulli (Boolean 0/1) random variable X with $p(X = 1) = q$ is given by

$$B(q) = -q \log q - (1 - q) \log(1 - q).$$

Suppose that a set S of examples contains p positive examples and n negative examples. The entropy of S is defined as $H(S) = B\left(\frac{p}{p+n}\right)$.

Based on an attribute X_j, we split our examples into k disjoint subsets S_k, with p_k positive and n_k negative examples in each. If the ratio $\frac{p_k}{p_k+n_k}$ is the same for all k, show that the information gain of this attribute is 0. **[5pts]**

Problem 3 (k-Nearest Neighbors)

In the following questions you will consider a k-nearest neighbor classifier using Euclidean distance metric on a binary classification task. We assign the class of the test point to be the class of the majority of the k nearest neighbors.

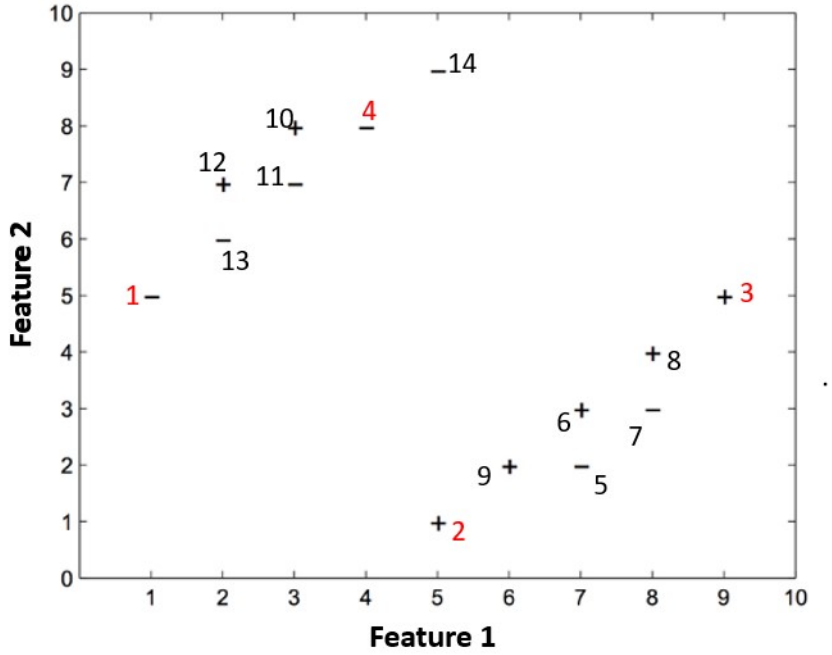

Figure 1: Dataset for KNN binary classification task.

(a) What value of k minimizes the training set error for this dataset? What is the resulting training error? (Assume that a point can be its own neighbor.) **[2pts]**

(b) What is the training error for the data points marked as $1, 2, 3, 4$ in Figure 1. Clearly mention which points are misclassified. (Assume that a point is not its own neighbour). **[4pts]**

In the above parts, we had considered k-nearest neighbour classifier using Euclidean distance metric. For the next part, we will choose *Cosine Similarity* as a distance metric. We define cosine similarity between two vectors $\mathbf{u}, \mathbf{v}$ as:

$$sim(\mathbf{u}, \mathbf{v}) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{||\mathbf{u}||_2 ||\mathbf{v}||_2}$$

where $< \cdot, \cdot >$ is the inner product operator between two vectors, and $||.||_2$ is the L_2 norm operator.

(c) What is the training error for the data points marked as $1, 2, 3, 4$ in Figure 1. Clearly mention which points are misclassified. (Assume that a point is not its own neighbour). **[4pts]**

Problem 4 (Kernel Properties)

(a) We know that the following is a valid Kernel: $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T\mathbf{z})^d$ where $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$, $d \in \mathbb{Z}^+$. Prove that $K_{new}(\mathbf{x}, \mathbf{z}) = \left(\sum_{i=1}^{n} \sqrt{x_i}\sqrt{z_i}\right)^d$ is a valid kernel.
Hint: If required you can prove and use the more general claim.
If $K(\mathbf{x}, \mathbf{z})$ is a valid kernel and $g(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^n$ then $K_{new}(\mathbf{x}, \mathbf{z}) = K(g(\mathbf{x}), g(\mathbf{z}))$ is also a valid kernel. **[5pts]**

(b) Is the following function K_2 a valid kernel?

$$K_2(\mathbf{x}, \mathbf{z}) = \frac{K_0(\mathbf{x}, \mathbf{z})}{\sqrt{K_0(\mathbf{x}, \mathbf{x})}\sqrt{K_0(\mathbf{z}, \mathbf{z})}},$$

where K_0 is a valid kernel. **[5pts]**

(c) Let $\phi : \mathbb{R}^n \to \mathbb{R}^m$, then under what conditions on the function ϕ, the following function $K_3(\mathbf{x}, \mathbf{z}) = K_0(\phi(\mathbf{x}), \phi(\mathbf{z}))$ is a valid kernel? Here, K_0 is a valid kernel in $\mathbb{R}^m$. **[5pts]**

Problem 5 (Bias-Variance Tradeoff)

In this problem, we ask you to compare the classification models we have studied till now i.e., logistic regression, KNN (n = 5), and Decision trees in terms of the high bias, high variance or balanced characteristics showcased by their learning curves on the Breast Cancer Dataset. **[20pts]**

Starter Notebook

---

code

- `Bias-Variance-Tradeoff.ipynb`

data

- Breast Cancer Dataset:
  https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html#sklearn.datasets.load_breast_cancer

documentation

- Decision Tree Classifier:
  http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- K-Nearest Neighbor Classifier:
  http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- Cross-Validation:
  http://scikit-learn.org/stable/modules/generated/sklearn.cross_validation.train_test_split.html
- Metrics:
  http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
- Learning Curves:
  https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.learning_curve.html

Learning curve is a plot which represents the trend of the training error and validation error with changing training set sizes averaged over multiple runs (cross-validated). Here, you are supposed to plot the learning curves for each model, and mention your inferences. The notebook provides you with the skeleton code to perform this task.

(a) Before we apply any machine learning problem, it is important for you to get comfortable with the dataset. Use the import functionality `load_breast_cancer()` to load the data in your notebook. `.data` and `.target` will provide you with the features, `X`, and their corresponding class labels, `y`. Once you do that, you will notice that the dataset has 569 examples, and each example has 30 features.

(b) Once you are done loading the data, we are going to define the models for logistic regression, K-nearest neighbor (K = 3) and Decision Tree one by one using `sklearn`. Make use of the documentation resources to define the `model` variable given in the code.

(c) The `learning_curve` method takes the `model`, `X`, `y`, `cv(default=5)`, and `train_sizes (default=[0.1, 0.3, 0.5, 0.7, 0.9])` as its input, and outputs `train_size`, `train_score` and `cross_validation_score`. `cv` indicates the number of cross-validation folds you want to run your model for. `train_sizes` indicates relative numbers of training examples that will be used to generate the learning curve. For instance, if we have 100 examples in our data, then with `cv=5` and `train_sizes=[0.1, 0.3])`, the model is going to return the training and validation scores for 5 different splits of data, where the size of the training split will be 10% in one case and 30% in the other. Finally, we average our results over different splits using `np.mean`. This will eventually leave you with the scores of your model across different training size splits.

(d) The `scores` returned by the `learning_curve` function are *accuracy* scores. We want to plot *errors* in this problem. Use the mathematical relationship between *error* and *accuracy* to get error scores.

(e) Finally, use the `matplotlib.pyplot` to plot the learning curves of all the models.

Problem 6 (Programming exercise: Applying decision trees and k-nearest neighbors)

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this problem, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

Starter Files

---

code and data

- code : `titanic.py`
- data : `titanic_train.csv`

Download the code and data sets from the course website. For more information on the data set, see the Kaggle description: https://www.kaggle.com/c/titanic/data. (The provided data sets are modified versions of the data available from Kaggle.[1])

Note that any portions of the code that you must modify have been indicated with `TODO`. Do not change any code outside of these blocks.

Visualization

One of the first things to do before trying any formal machine learning technique is to dive into the data. This can include looking for funny values in the data, looking for outliers, looking at the range of feature values, what features seem important, etc.

(a) Run the code (`titanic.py`) to make histograms for each feature, separating the examples by class (e.g. survival). This should produce seven plots, one for each feature, and each plot should have two overlapping histograms, with the color of the histogram indicating the class. For each feature, what trends do you observe in the data? **[5pts]**

Evaluation

Now, let us use `scikit-learn` to train a `DecisionTreeClassifier` and `KNeighborsClassifier` on the data.

Using the predictive capabilities of the `scikit-learn` package is very simple. In fact, it can be carried out in three simple steps: initializing the model, fitting it to the training data, and predicting new values.[2]

(b) Before trying out any classifier, it is often useful to establish a *baseline*. We have implemented one simple baseline classifier, `MajorityVoteClassifier`, that always predicts the majority class from the training set. Read through the `MajorityVoteClassifier` and its usage and make sure you understand how it works.

[1] Passengers with missing values for any feature have been removed. Also, the categorical feature `Sex` has been mapped to `{'female': 0, 'male': 1}` and `Embarked` to `{'C': 0, 'Q': 1, 'S': 2}`. If you are interested more in this process of *data munging*, Kaggle has an excellent tutorial available at https://www.kaggle.com/c/titanic/details/getting-started-with-python-ii.

[2] Note that almost all of the model techniques in `scikit-learn` share a few common named functions, once they are initialized. You can always find out more about them in the documentation for each model. These are `some-model-name.fit(...)`, `some-model-name.predict(...)`, and `some-model-name.score(...)`.

Your goal is to implement and evaluate another baseline classifier, `RandomClassifier`, that predicts a target class according to the distribution of classes in the training data set. For example, if 60% of the examples in the training set have `Survived = 0` and 40% have `Survived = 1`, then, when applied to a test set, `RandomClassifier` should randomly predict 60% of the examples as `Survived = 0` and 40% as `Survived = 1`.

Implement the missing portions of `RandomClassifier` according to the provided specifications. Then train your `RandomClassifier` on the entire training data set, and evaluate its training error. If you implemented everything correctly, you should have an error of 0.485.

(c) Now that we have a baseline, train and evaluate a `DecisionTreeClassifier` (using the class from `scikit-learn` and referring to the documentation as needed). Make sure you initialize your classifier with the appropriate parameters; in particular, use the 'entropy' criterion discussed in class. What is the training error of this classifier? **[5pts]**

(d) Similar to the previous question, train and evaluate a `KNeighborsClassifier` (using the class from `scikit-learn` and referring to the documentation as needed). Use k=3, 5 and 7 as the number of neighbors and report the training error of this classifier. **[5pts]**

(e) So far, we have looked only at training error, but as we learned in class, training error is a poor metric for evaluating classifiers. Let us use cross-validation instead.

Implement the missing portions of `error(...)` according to the provided specifications. You may find it helpful to use `train_test_split(...)` from `scikit-learn`. To ensure that we always get the same splits across different runs (and thus can compare the classifier results), set the `random_state` parameter to be the trial number.

Next, use your `error(...)` function to evaluate the training error and (cross-validation) test error of each of your four models (for the `KNeighborsClassifier`, use k=5). To do this, generate a random 80/20 split of the training data, train each model on the 80% fraction, evaluate the error on either the 80% or the 20% fraction, and repeat this 100 times to get an average result. What are the average training and test error of each of your classifiers on the Titanic data set? **[10pts]**

(f) One way to find out the best value of k for `KNeighborsClassifier` is n-fold cross validation. Find out the best value of k using 10-fold cross validation. You may find the `cross_val_score(...)` from `scikit-learn` helpful. Run 10-fold cross validation for all odd numbers ranging from 1 to 50 as the number of neighbors. Then plot the validation error against the number of neighbors, k. Include this plot in your writeup, and provide a 1-2 sentence description of your observations. What is the best value of k? **[5pts]**

(g) One problem with decision trees is that they can *overfit* to training data, yielding complex classifiers that do not generalize well to new data. Let us see whether this is the case for the Titanic data.

One way to prevent decision trees from overfitting is to limit their depth. Repeat your cross-validation experiments but for increasing depth limits, specifically, $1, 2, \ldots, 20$. Then plot the average training error and test error against the depth limit. Include this plot in your writeup, making sure to label all axes and include a legend for your classifiers. What is the best depth limit to use for this data? Do you see overfitting? Justify your answers using the plot. **[5pts]**

(h) Another useful tool for evaluating classifiers is *learning curves*, which show how classifier performance (e.g. error) relates to experience (e.g. amount of training data). For this experiment, first generate a random 90/10 split of the training data and do the following experiments considering the 90% fraction as training and 10% for testing.

Run experiments for the decision tree and k-nearest neighbors classifier with the best depth limit and k value you found above. This time, vary the amount of training data by starting with splits of 0.10 (10% of the data from 90% fraction) and working up to full size 1.00 (100% of the data from 90% fraction) in increments of 0.10. Then plot the decision tree and k-nearest neighbors training and test error against the amount of training data. Include this plot in your writeup, and provide a 1-2 sentence description of your observations. **[10pts]**