CS6316: HW3

Aobo Yang (ay6gv)

October 22, 2019

1. KNN and Model Selection (k)

   1.6

   The best k is 7 and the corresponding accuracies are shown in the table below. The reason of that some k works better than others is that k decides the model complexity. Smaller k may make the model too complicate and easier to be affected by noises nearby, so it overfits the training set. Larger k, on the other hand, may make the model too generic, so it underfits.

K	Accuracy
3	0.6155
5	0.6275
7	0.629
9	0.626
11	0.6285
13	0.6255

   1.7

   The bar graph between k and accuracy is shown in 1

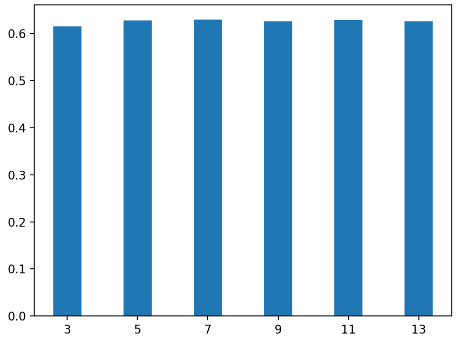

Figure 1: KNN Bar

2. Support Vector Machines

   The 3-fold cross-validation accuracies of different hyperparameters are shown in the table below.

kernel	C	degree	training accuracy	validation accuracy
linear	1		0.8522	0.8514
linear	10		0.8523	0.8516
linear	100		0.8523	0.8516
rbf	1		0.8541	0.8519
rbf	10		0.8612	0.8546
rbf	100		0.8712	0.8563
rbf	1000		0.8875	0.8495
poly	1	1	0.8504	0.8508
poly	1	3	0.8207	0.8196
poly	1	5	0.7778	0.7775
poly	10	1	0.8521	0.8511
poly	10	3	0.8459	0.8425
poly	10	5	0.7863	0.7849

The best performing model is the one with the "rbf" kernel and C value of 100. It achieves the highest validation accuracy 0.8563.

The data preprocessing contains three steps. First, I use LabelEncoder to map the target labels to 0 and 1. Second, I use scikit-learn's StandardScaler to normalize all the continuous attributes by removing the mean and scaling to unit variance. At last, I use scikit-learn's OneHotEncoder to expand all the categorical features except "native-country". I decide to drop the "native-country" because it alone brings in around 40 new one-hot features which dramatically hurts the SVM training speed and I find having it does not contribute much to the prediction.

3. Sample QA Questions

   (a)

   False, larger C penalize violations more so there should be less data fall in the smaller margin which means less support vectors. On the contrary, smaller C leads to larger margin so there are more support vectors.

   (b)

   Correct option (1)

   (c)

   (2) (1) (3)