sms-spam-v3

April 23, 2019

V3:

- LSTM uses the Tokenizert.fit_on_texts(data) then Tokenizert.texts_to_sequences(data) to do the preprocessing
- Other models uses the TfidfVectorizer to do the the preprocessing

Goal of this notebook to test several classifiers on the data set with different features

0.0.1 Let's begin

First of all neccesary imports

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib as mpl
        import matplotlib.pyplot as plt
        import seaborn as sns
        import string
        from sklearn.feature_extraction.text import TfidfVectorizer
        from sklearn.model_selection import train_test_split
        from nltk.corpus import stopwords
        from sklearn.preprocessing import LabelEncoder
        %matplotlib inline
```

Let's read the data from csv file

```
In [2]: sms = pd.read_csv('../data/sms-spam.csv',delimiter=',',encoding='latin-1')

        sms.head()
```

```
Out[2]:      v1                                                 v2 Unnamed: 2  \
        0   ham  Go until jurong point, crazy.. Available only ...        NaN
        1   ham                      Ok lar... Joking wif u oni...        NaN
        2  spam  Free entry in 2 a wkly comp to win FA Cup fina...        NaN
        3   ham  U dun say so early hor... U c already then say...        NaN
        4   ham  Nah I don't think he goes to usf, he lives aro...        NaN

          Unnamed: 3 Unnamed: 4
```

```
0        NaN        NaN
1        NaN        NaN
2        NaN        NaN
3        NaN        NaN
4        NaN        NaN
```

Now drop "unnamed" columns and rename v1 and v2 to "label" and "message"

```
In [3]: sms = sms.drop(['Unnamed: 2','Unnamed: 3','Unnamed: 4'],axis=1)
        sms = sms.rename(columns = {'v1':'label','v2':'message'})
```

Let's look into our data

```
In [4]: sms.groupby('label').describe()
```

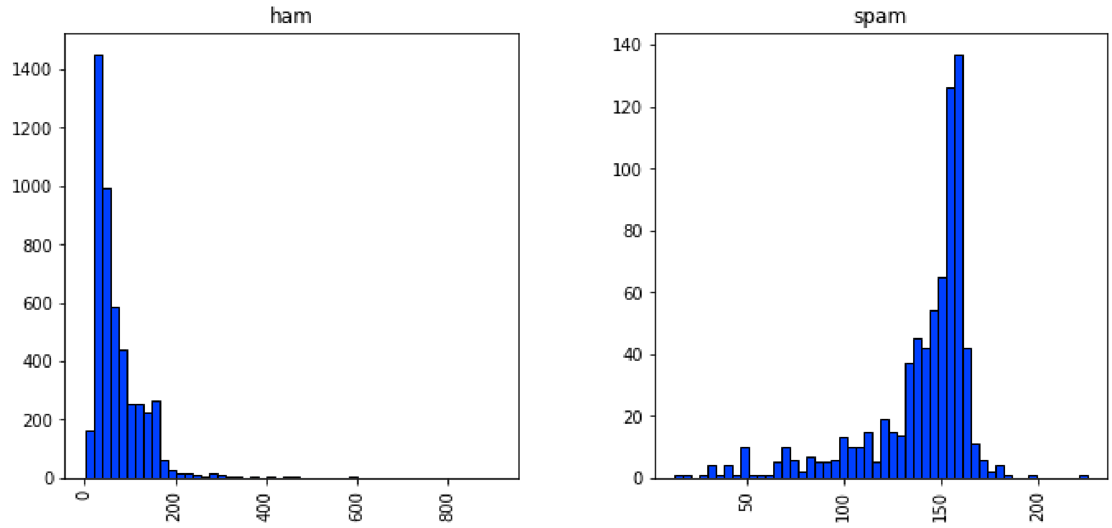
```
Out[4]:      message
              count unique                                              top freq
        label
        ham    4825   4516                          Sorry, I'll call later   30
        spam    747    653  Please call our customer service representativ...    4
```

Intresting that "Sorry, I'll call later" appears only 30 times here =)
Now let's create new feature "message length" and plot it to see if it's of any interest

```
In [5]: sms['length'] = sms['message'].apply(len)
        sms.head()
```

```
Out[5]:   label                                            message  length
        0   ham  Go until jurong point, crazy.. Available only ...     111
        1   ham                      Ok lar... Joking wif u oni...      29
        2  spam  Free entry in 2 a wkly comp to win FA Cup fina...     155
        3   ham  U dun say so early hor... U c already then say...      49
        4   ham  Nah I don't think he goes to usf, he lives aro...      61
```

```
In [6]: mpl.rcParams['patch.force_edgecolor'] = True
        plt.style.use('seaborn-bright')
        sms.hist(column='length', by='label', bins=50,figsize=(11,5))
        plt.savefig("../img/sms-length-distribution.eps")
        plt.show()
```



Looks like the lengthy is the message, more likely it is a spam. Let's not forget this

0.0.2 Text processing and vectorizing our meddages

Let's create new data frame. We'll need a copy later on

```
In [7]: text_feat = sms['message'].copy()
```

Now define our tex precessing function. It will remove any punctuation and stopwords aswell.

```
In [8]: # def text_process(text):

        #     text = text.translate(str.maketrans('', '', string.punctuation))
        #     text = [word for word in text.split() if word.lower() not in stopwords.words('en

        #     return " ".join(text)

In [9]: # text_feat = text_feat.apply(text_process)

In [10]: vectorizer = TfidfVectorizer("english")

In [11]: features = vectorizer.fit_transform(text_feat)

In [12]: labels = LabelEncoder().fit_transform(sms['label'])
         labels = labels.reshape(-1,1)

In [13]: text_feat.shape

Out[13]: (5572,)

In [14]: features.shape

Out[14]: (5572, 8710)
```

0.0.3 Classifiers and predictions

First of all let's split our features to test and train set

```
In [15]: features_train, features_test, labels_train, labels_test = train_test_split(features,
```

Now let's import bunch of classifiers, initialize them and make a dictionary to itereate through

```
In [16]: from sklearn.linear_model import LogisticRegression
         from sklearn.svm import SVC
         from sklearn.naive_bayes import MultinomialNB
         from sklearn.tree import DecisionTreeClassifier
         from sklearn.neighbors import KNeighborsClassifier
         from sklearn.ensemble import RandomForestClassifier
         from sklearn.ensemble import AdaBoostClassifier
         from sklearn.ensemble import BaggingClassifier
         from sklearn.ensemble import ExtraTreesClassifier
         from sklearn.metrics import precision_score, recall_score, accuracy_score, f1_score
```

```
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/ensemble/weight_boosting.py:29
  from numpy.core.umath_tests import inner1d
```

```
In [17]: svc = SVC(kernel='sigmoid', gamma=1.0)
         knc = KNeighborsClassifier(n_neighbors=49)
         mnb = MultinomialNB(alpha=0.2)
         dtc = DecisionTreeClassifier(min_samples_split=7, random_state=111)
         lrc = LogisticRegression(solver='liblinear', penalty='l1')
         rfc = RandomForestClassifier(n_estimators=31, random_state=111)
         abc = AdaBoostClassifier(n_estimators=62, random_state=111)
         bc = BaggingClassifier(n_estimators=9, random_state=111)
         etc = ExtraTreesClassifier(n_estimators=9, random_state=111)
```

```
In [18]: clfs = {'SVC' : svc,'KN' : knc, 'NB': mnb, 'DT': dtc, 'LR': lrc, 'RF': rfc, 'AdaBoost
```

Let's make functions to fit our classifiers and make predictions

```
In [19]: def train_classifier(clf, feature_train, labels_train):
             clf.fit(feature_train, labels_train)
```

```
In [20]: def predict_labels(clf, features):
             return (clf.predict(features))
```

Now iterate through classifiers and save the results

```
In [21]: import time
```

```
In [22]: pred_scores = []
         for k,v in clfs.items():
             since = time.time()
```

```
            train_classifier(v, features_train, labels_train)
            time_elapsed = time.time() - since

            pred = predict_labels(v,features_test)
            pred_scores.append((k, [precision_score(labels_test,pred), recall_score(labels_tes
```

```
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/utils/validation.py:578: Data
  y = column_or_1d(y, warn=True)
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/ipykernel/__main__.py:2: DataConversi
  from ipykernel import kernelapp as app
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/ipykernel/__main__.py:2: DataConversi
  from ipykernel import kernelapp as app
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/utils/validation.py:578: Data
  y = column_or_1d(y, warn=True)
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/utils/validation.py:578: Data
  y = column_or_1d(y, warn=True)
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/ipykernel/__main__.py:2: DataConversi
  from ipykernel import kernelapp as app
```

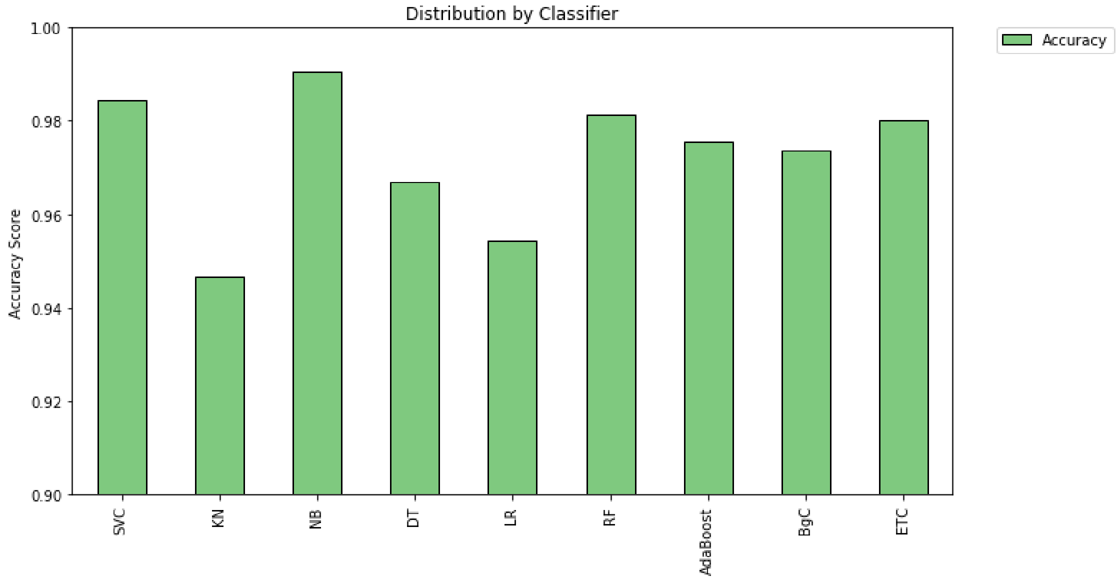
In [23]:
```
# pred_scores
```

In [24]:
```
df = pd.DataFrame.from_items(pred_scores,orient='index', columns=['Precision', 'Recal
df
```

Out[24]:

	Precision	Recall	Accuracy	F1	Training Time (s)
SVC	0.990476	0.896552	0.984450	0.941176	0m 0.4593s
KN	1.000000	0.616379	0.946770	0.762667	0m 0.0041s
NB	0.982143	0.948276	0.990431	0.964912	0m 0.0018s
DT	0.900452	0.857759	0.967105	0.878587	0m 0.2270s
LR	0.933333	0.724138	0.954545	0.815534	0m 0.0107s
RF	1.000000	0.866379	0.981459	0.928406	0m 0.9180s
AdaBoost	0.952607	0.866379	0.975478	0.907449	0m 2.4091s
BgC	0.935185	0.870690	0.973684	0.901786	0m 1.1530s
ETC	0.995025	0.862069	0.980263	0.923788	0m 0.6672s

In [25]:
```
df.plot(kind='bar', y="Accuracy", ylim=(0.9,1.0), figsize=(11,6), align='center', colo
plt.xticks(np.arange(9), df.index)
plt.ylabel('Accuracy Score')
plt.title('Distribution by Classifier')
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
plt.savefig("../img/sms-acc-basemodel-v3.eps")
plt.show()
```



Looks like ensemble classifiers are not doing as good as expected.

0.0.4 Voting classifier

We are using ensemble algorithms here, but what about ensemble of ensembles? Will it beat NB?

```
In [26]: from sklearn.ensemble import VotingClassifier

In [27]: eclf = VotingClassifier(estimators=[('BgC', bc), ('ETC', etc), ('RF', rfc), ('Ada', al

In [28]: eclf.fit(features_train,labels_train)

/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/preprocessing/label.py:95: Da
  y = column_or_1d(y, warn=True)
/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/preprocessing/label.py:128: Da
  y = column_or_1d(y, warn=True)


Out[28]: VotingClassifier(estimators=[('BgC', BaggingClassifier(base_estimator=None, bootstrap=
                  bootstrap_features=False, max_features=1.0, max_samples=1.0,
                  n_estimators=9, n_jobs=1, oob_score=False, random_state=111,
                  verbose=0, warm_start=False)), ('ETC', ExtraTreesClassifier(bootstrap=False,
                   learning_rate=1.0, n_estimators=62, random_state=111))],
                  flatten_transform=None, n_jobs=1, voting='soft', weights=None)

In [29]: pred = eclf.predict(features_test)

/Users/alex/anaconda/envs/gc/lib/python3.6/site-packages/sklearn/preprocessing/label.py:151: De
  if diff:
```

```
In [30]: print(precision_score(labels_test,pred), recall_score(labels_test,pred), accuracy_sco
```

```
0.9806763285024155 0.875 0.9802631578947368 0.9248291571753987
```

Better but nope.

0.0.5 RNN

Define the RNN structure.

```
In [31]: from keras.models import Model
         from keras.layers import LSTM, Activation, Dense, Dropout, Input, Embedding
         from keras.optimizers import RMSprop
         from keras.preprocessing.text import Tokenizer
         from keras.preprocessing import sequence
         from keras.utils import to_categorical
         from keras.callbacks import EarlyStopping
         from keras.callbacks import Callback
```

```
Using TensorFlow backend.
```

0.0.6 Process the data

- Tokenize the data and convert the text to sequences.
- Add padding to ensure that all the sequences have the same shape.
- There are many ways of taking the *max_len* and here an arbitrary length of 500 is chosen. (From the Fig, almost all the sentences have the length < 200)

```
In [32]: features_lstm = text_feat
         labels_lstm = labels

In [33]: max_words = 1000
         max_len = 200 # n_features
         tok = Tokenizer(num_words=max_words)
         tok.fit_on_texts(features_lstm)
         sequences = tok.texts_to_sequences(features_lstm)
         features_lstm = sequence.pad_sequences(sequences,maxlen=max_len)

In [34]: features_lstm.shape

Out[34]: (5572, 200)

In [35]: labels_lstm.shape

Out[35]: (5572, 1)

In [36]: features_lstm_train, features_lstm_test, labels_lstm_train, labels_lstm_test = train_
```

```
In [37]: def RNN():
             inputs = Input(name='inputs',shape=[max_len])
             layer = Embedding(max_words,50,input_length=max_len)(inputs)
             layer = LSTM(64)(layer)
             layer = Dense(256,name='FC1')(layer)
             layer = Activation('relu')(layer)
             layer = Dropout(0.5)(layer)
             layer = Dense(1,name='out_layer')(layer)
             layer = Activation('sigmoid')(layer)
             model = Model(inputs=inputs,outputs=layer)
             return model
```

Call the function and compile the model.

```
In [38]: model = RNN()
         model.summary()
         model.compile(loss='binary_crossentropy',optimizer=RMSprop(),metrics=['accuracy'])
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
inputs (InputLayer)          (None, 200)               0
_________________________________________________________________
embedding_1 (Embedding)      (None, 200, 50)           50000
_________________________________________________________________
lstm_1 (LSTM)                (None, 64)                29440
_________________________________________________________________
FC1 (Dense)                  (None, 256)               16640
_________________________________________________________________
activation_1 (Activation)    (None, 256)               0
_________________________________________________________________
dropout_1 (Dropout)          (None, 256)               0
_________________________________________________________________
out_layer (Dense)            (None, 1)                 257
_________________________________________________________________
activation_2 (Activation)    (None, 1)                 0
=================================================================
Total params: 96,337
Trainable params: 96,337
Non-trainable params: 0
_________________________________________________________________
```

```
In [39]: since = time.time()


         model.fit(features_lstm_train, labels_lstm_train, epochs=10, batch_size=128,validatio
                               callbacks=[EarlyStopping(monitor='val_loss',min_delta=0.0001)])
```

```
time_elapsed = time.time() - since
```

```
Train on 3120 samples, validate on 780 samples
Epoch 1/10
3120/3120 [==============================] - 9s 3ms/step - loss: 0.3671 - acc: 0.8538 - val_los
Epoch 2/10
3120/3120 [==============================] - 8s 3ms/step - loss: 0.1145 - acc: 0.9696 - val_los
Epoch 3/10
3120/3120 [==============================] - 6s 2ms/step - loss: 0.0458 - acc: 0.9869 - val_los
```

```
In [40]: print('Training complete in {:.0f}m {:.4f}s'.format(
                 time_elapsed // 60, time_elapsed % 60))
```

```
Training complete in 0m 23.4645s
```

```
In [41]: pred = (np.asarray(model.predict(features_lstm_test, batch_size=128))).round()
```

```
In [42]: pred_scores.append(("LSTM", [precision_score(labels_lstm_test,pred), recall_score(labe
```

0.0.7 gcForest

```
In [43]: import sys
         sys.path.append("..")
         from gcforest.gcforest import GCForest
         from gcforest.utils.config_utils import load_json
```

```
In [44]: def get_toy_config():
             config = {}
             ca_config = {}
             ca_config["random_state"] = 111
             ca_config["max_layers"] = 10
             ca_config["early_stopping_rounds"] = 3
             ca_config["n_classes"] = 2
             ca_config["estimators"] = []
             ca_config["estimators"].append({"n_folds": 5, "type": "RandomForestClassifier", "r
             ca_config["estimators"].append({"n_folds": 5, "type": "MultinomialNB", "alpha": 0

             config["cascade"] = ca_config
             return config
```

```
In [45]: config = get_toy_config()
         gc = GCForest(config)

         # features_train  ndarraylabels_train  (n_samples, )(n_samples, 1)
         features_train = features_train.toarray()
         labels_train = labels_train.reshape(-1)
```

```python
since = time.time()
gc.fit_transform(features_train, labels_train)

time_elapsed = time.time() - since

# gc.fit_transform(features_train, labels_train, features_test, labels_test)
```

```
[ 2019-04-23 22:00:22,133][cascade_classifier.fit_transform] X_groups_train.shape=[(3900, 8710)
[ 2019-04-23 22:00:22,356][cascade_classifier.fit_transform] group_dims=[8710]
[ 2019-04-23 22:00:22,358][cascade_classifier.fit_transform] group_starts=[0]
[ 2019-04-23 22:00:22,358][cascade_classifier.fit_transform] group_ends=[8710]
[ 2019-04-23 22:00:22,359][cascade_classifier.fit_transform] X_train.shape=(3900, 8710),X_test
[ 2019-04-23 22:00:22,527][cascade_classifier.fit_transform] [layer=0] look_indexs=[0], X_cur_
[ 2019-04-23 22:00:25,099][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 22:00:27,622][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 22:00:30,116][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 22:00:32,607][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 22:00:35,006][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 22:00:35,007][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_0 - 5_
[ 2019-04-23 22:00:35,218][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 22:00:35,427][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 22:00:35,634][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 22:00:35,844][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 22:00:36,051][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 22:00:36,052][kfold_wrapper.log_eval_metrics] Accuracy(layer_0 - estimator_1 - 5_
[ 2019-04-23 22:00:36,053][cascade_classifier.calc_accuracy] Accuracy(layer_0 - train.classifi
[ 2019-04-23 22:00:36,227][cascade_classifier.fit_transform] [layer=1] look_indexs=[0], X_cur_
[ 2019-04-23 22:00:38,326][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 22:00:40,314][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 22:00:42,200][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 22:00:44,191][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 22:00:45,962][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 22:00:45,964][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_0 - 5_
[ 2019-04-23 22:00:46,181][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 22:00:46,397][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 22:00:46,614][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 22:00:46,832][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 22:00:47,049][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 22:00:47,050][kfold_wrapper.log_eval_metrics] Accuracy(layer_1 - estimator_1 - 5_
[ 2019-04-23 22:00:47,052][cascade_classifier.calc_accuracy] Accuracy(layer_1 - train.classifi
[ 2019-04-23 22:00:47,240][cascade_classifier.fit_transform] [layer=2] look_indexs=[0], X_cur_
[ 2019-04-23 22:00:48,928][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 22:00:50,611][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 22:00:52,392][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 22:00:54,081][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 22:00:55,766][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
[ 2019-04-23 22:00:55,767][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_0 - 5_
```

```
[ 2019-04-23 22:00:55,988][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 22:00:56,206][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 22:00:56,424][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 22:00:56,642][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 22:00:56,859][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 22:00:56,860][kfold_wrapper.log_eval_metrics] Accuracy(layer_2 - estimator_1 - 5_
[ 2019-04-23 22:00:56,861][cascade_classifier.calc_accuracy] Accuracy(layer_2 - train.classifi
[ 2019-04-23 22:00:57,031][cascade_classifier.fit_transform] [layer=3] look_indexs=[0], X_cur_
[ 2019-04-23 22:00:58,823][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 22:01:00,703][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 22:01:02,374][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 22:01:04,166][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 22:01:05,834][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 22:01:05,835][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_0 - 5_
[ 2019-04-23 22:01:06,055][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 22:01:06,276][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 22:01:06,494][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 22:01:06,719][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 22:01:06,935][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 22:01:06,936][kfold_wrapper.log_eval_metrics] Accuracy(layer_3 - estimator_1 - 5_
[ 2019-04-23 22:01:06,937][cascade_classifier.calc_accuracy] Accuracy(layer_3 - train.classifi
[ 2019-04-23 22:01:07,166][cascade_classifier.fit_transform] [layer=4] look_indexs=[0], X_cur_
[ 2019-04-23 22:01:08,761][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 22:01:10,336][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 22:01:12,021][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 22:01:13,590][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 22:01:15,261][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 22:01:15,262][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_0 - 5_
[ 2019-04-23 22:01:15,472][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 22:01:15,680][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 22:01:15,893][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 22:01:16,103][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 22:01:16,314][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 22:01:16,315][kfold_wrapper.log_eval_metrics] Accuracy(layer_4 - estimator_1 - 5_
[ 2019-04-23 22:01:16,316][cascade_classifier.calc_accuracy] Accuracy(layer_4 - train.classifi
[ 2019-04-23 22:01:16,489][cascade_classifier.fit_transform] [layer=5] look_indexs=[0], X_cur_
[ 2019-04-23 22:01:18,274][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_0 - 5_
[ 2019-04-23 22:01:19,942][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_0 - 5_
[ 2019-04-23 22:01:21,401][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_0 - 5_
[ 2019-04-23 22:01:22,966][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_0 - 5_
[ 2019-04-23 22:01:24,641][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_0 - 5_
[ 2019-04-23 22:01:24,643][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_0 - 5_
[ 2019-04-23 22:01:24,847][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_1 - 5_
[ 2019-04-23 22:01:25,052][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_1 - 5_
[ 2019-04-23 22:01:25,259][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_1 - 5_
[ 2019-04-23 22:01:25,466][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_1 - 5_
[ 2019-04-23 22:01:25,670][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_1 - 5_
[ 2019-04-23 22:01:25,671][kfold_wrapper.log_eval_metrics] Accuracy(layer_5 - estimator_1 - 5_
```

```
[ 2019-04-23 22:01:25,672][cascade_classifier.calc_accuracy] Accuracy(layer_5 - train.classifi
[ 2019-04-23 22:01:25,845][cascade_classifier.fit_transform] [layer=6] look_indexs=[0], X_cur_
[ 2019-04-23 22:01:27,616][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_0 - 5_
[ 2019-04-23 22:01:29,187][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_0 - 5_
[ 2019-04-23 22:01:30,861][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_0 - 5_
[ 2019-04-23 22:01:32,611][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_0 - 5_
[ 2019-04-23 22:01:34,286][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_0 - 5_
[ 2019-04-23 22:01:34,288][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_0 - 5_
[ 2019-04-23 22:01:34,499][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_1 - 5_
[ 2019-04-23 22:01:34,710][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_1 - 5_
[ 2019-04-23 22:01:34,917][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_1 - 5_
[ 2019-04-23 22:01:35,129][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_1 - 5_
[ 2019-04-23 22:01:35,337][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_1 - 5_
[ 2019-04-23 22:01:35,339][kfold_wrapper.log_eval_metrics] Accuracy(layer_6 - estimator_1 - 5_
[ 2019-04-23 22:01:35,340][cascade_classifier.calc_accuracy] Accuracy(layer_6 - train.classifi
[ 2019-04-23 22:01:35,514][cascade_classifier.fit_transform] [layer=7] look_indexs=[0], X_cur_
[ 2019-04-23 22:01:37,199][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_0 - 5_
[ 2019-04-23 22:01:38,973][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_0 - 5_
[ 2019-04-23 22:01:40,633][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_0 - 5_
[ 2019-04-23 22:01:42,194][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_0 - 5_
[ 2019-04-23 22:01:43,769][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_0 - 5_
[ 2019-04-23 22:01:43,771][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_0 - 5_
[ 2019-04-23 22:01:43,983][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_1 - 5_
[ 2019-04-23 22:01:44,189][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_1 - 5_
[ 2019-04-23 22:01:44,397][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_1 - 5_
[ 2019-04-23 22:01:44,604][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_1 - 5_
[ 2019-04-23 22:01:44,812][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_1 - 5_
[ 2019-04-23 22:01:44,814][kfold_wrapper.log_eval_metrics] Accuracy(layer_7 - estimator_1 - 5_
[ 2019-04-23 22:01:44,815][cascade_classifier.calc_accuracy] Accuracy(layer_7 - train.classifi
[ 2019-04-23 22:01:44,982][cascade_classifier.fit_transform] [layer=8] look_indexs=[0], X_cur_
[ 2019-04-23 22:01:46,650][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_0 - 5_
[ 2019-04-23 22:01:48,322][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_0 - 5_
[ 2019-04-23 22:01:49,992][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_0 - 5_
[ 2019-04-23 22:01:51,661][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_0 - 5_
[ 2019-04-23 22:01:53,338][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_0 - 5_
[ 2019-04-23 22:01:53,340][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_0 - 5_
[ 2019-04-23 22:01:53,547][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_1 - 5_
[ 2019-04-23 22:01:53,758][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_1 - 5_
[ 2019-04-23 22:01:53,969][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_1 - 5_
[ 2019-04-23 22:01:54,178][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_1 - 5_
[ 2019-04-23 22:01:54,388][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_1 - 5_
[ 2019-04-23 22:01:54,389][kfold_wrapper.log_eval_metrics] Accuracy(layer_8 - estimator_1 - 5_
[ 2019-04-23 22:01:54,390][cascade_classifier.calc_accuracy] Accuracy(layer_8 - train.classifi
[ 2019-04-23 22:01:54,562][cascade_classifier.fit_transform] [layer=9] look_indexs=[0], X_cur_
[ 2019-04-23 22:01:56,332][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_0 - 5_
[ 2019-04-23 22:01:57,896][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_0 - 5_
[ 2019-04-23 22:01:59,668][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_0 - 5_
[ 2019-04-23 22:02:01,237][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_0 - 5_
```

```
[ 2019-04-23 22:02:02,809][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_0 - 5_
[ 2019-04-23 22:02:02,813][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_0 - 5_
[ 2019-04-23 22:02:03,020][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_1 - 5_
[ 2019-04-23 22:02:03,227][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_1 - 5_
[ 2019-04-23 22:02:03,437][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_1 - 5_
[ 2019-04-23 22:02:03,643][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_1 - 5_
[ 2019-04-23 22:02:03,852][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_1 - 5_
[ 2019-04-23 22:02:03,853][kfold_wrapper.log_eval_metrics] Accuracy(layer_9 - estimator_1 - 5_
[ 2019-04-23 22:02:03,855][cascade_classifier.calc_accuracy] Accuracy(layer_9 - train.classifi
[ 2019-04-23 22:02:03,856][cascade_classifier.fit_transform] [Result][Reach Max Layer] opt_laye
```

In [46]:
```
print('Training complete in {:.0f}m {:.4f}s'.format(
                time_elapsed // 60, time_elapsed % 60))
```

```
Training complete in 1m 41.7678s
```

In [47]:
```
pred = predict_labels(gc,features_test.toarray())
pred_scores.append(("DCF", [precision_score(labels_test,pred), recall_score(labels_tes
```

```
[ 2019-04-23 22:02:03,955][cascade_classifier.transform] X_groups_test.shape=[(1672, 8710)]
[ 2019-04-23 22:02:04,064][cascade_classifier.transform] group_dims=[8710]
[ 2019-04-23 22:02:04,065][cascade_classifier.transform] X_test.shape=(1672, 8710)
[ 2019-04-23 22:02:04,131][cascade_classifier.transform] [layer=0] look_indexs=[0], X_cur_test
[ 2019-04-23 22:02:04,983][cascade_classifier.transform] [layer=1] look_indexs=[0], X_cur_test
[ 2019-04-23 22:02:05,795][cascade_classifier.transform] [layer=2] look_indexs=[0], X_cur_test
[ 2019-04-23 22:02:06,605][cascade_classifier.transform] [layer=3] look_indexs=[0], X_cur_test
[ 2019-04-23 22:02:07,421][cascade_classifier.transform] [layer=4] look_indexs=[0], X_cur_test
[ 2019-04-23 22:02:08,230][cascade_classifier.transform] [layer=5] look_indexs=[0], X_cur_test
[ 2019-04-23 22:02:09,037][cascade_classifier.transform] [layer=6] look_indexs=[0], X_cur_test
[ 2019-04-23 22:02:09,847][cascade_classifier.transform] [layer=7] look_indexs=[0], X_cur_test
[ 2019-04-23 22:02:10,654][cascade_classifier.transform] [layer=8] look_indexs=[0], X_cur_test
[ 2019-04-23 22:02:11,456][cascade_classifier.transform] [layer=9] look_indexs=[0], X_cur_test
```

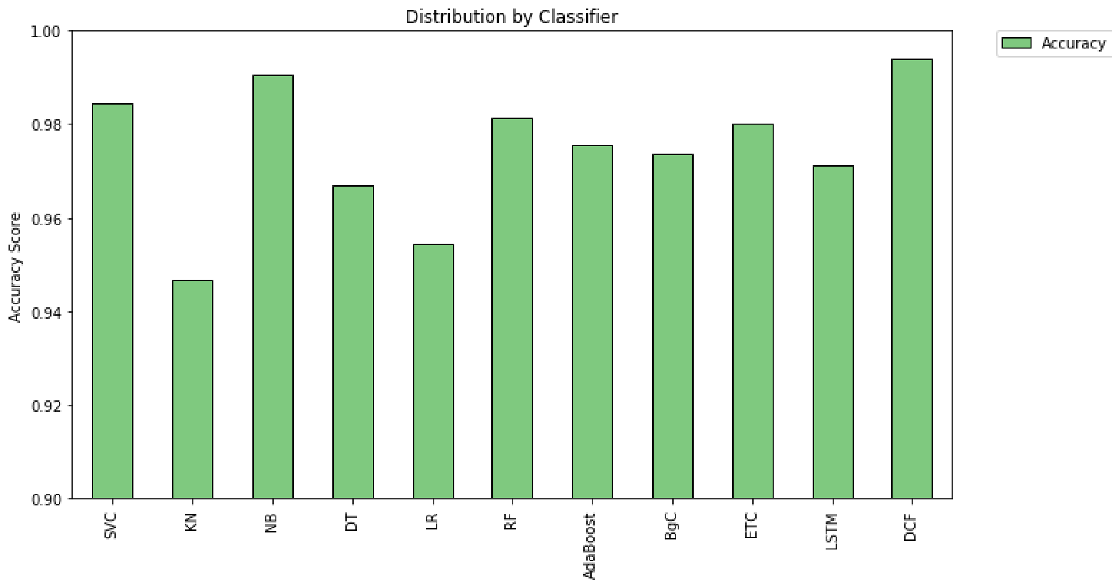
In [48]:
```
df = pd.DataFrame.from_items(pred_scores,orient='index', columns=['Precision', 'Recal
df
```

Out[48]:

	Precision	Recall	Accuracy	F1	Training Time (s)
SVC	0.990476	0.896552	0.984450	0.941176	0m 0.4593s
KN	1.000000	0.616379	0.946770	0.762667	0m 0.0041s
NB	0.982143	0.948276	0.990431	0.964912	0m 0.0018s
DT	0.900452	0.857759	0.967105	0.878587	0m 0.2270s
LR	0.933333	0.724138	0.954545	0.815534	0m 0.0107s
RF	1.000000	0.866379	0.981459	0.928406	0m 0.9180s
AdaBoost	0.952607	0.866379	0.975478	0.907449	0m 2.4091s
BgC	0.935185	0.870690	0.973684	0.901786	0m 1.1530s
ETC	0.995025	0.862069	0.980263	0.923788	0m 0.6672s

```
LSTM        0.854839  0.946429  0.971292  0.898305       0m 23.4645s
DCF         0.986842  0.969828  0.994019  0.978261       1m 41.7678s
```

In [49]:
```python
df.plot(kind='bar', y="Accuracy", ylim=(0.9,1.0), figsize=(11,6), align='center', colo
plt.xticks(np.arange(11), df.index)
plt.ylabel('Accuracy Score')
plt.title('Distribution by Classifier')
plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)
plt.savefig("../img/sms-acc-v3.eps")
plt.show()
```



In [50]:
```python
import pickle
# dump
with open("../pkl/sms-gc-v3.pkl", "wb") as f:
    pickle.dump(gc, f, pickle.HIGHEST_PROTOCOL)

# # load
# with open("../pkl/2018_gc.pkl", "rb") as f:
#     gc = pickle.load(f)
```

0.0.8 Final verdict - gcForest is your friend in spam detection.

In [ ]: