Applied Machine Learning

INFO 370

Learning Objectives

Review machine learning process

Learn how to use the Driven Data submission system

Predict blood donations!

Machine Learning Review

Machine Learning Steps

Import a classifier (a method that generates an instance of a classifier)

Create a classifier

Split train/test data

Find optimal parameters (grid search) through a process that:

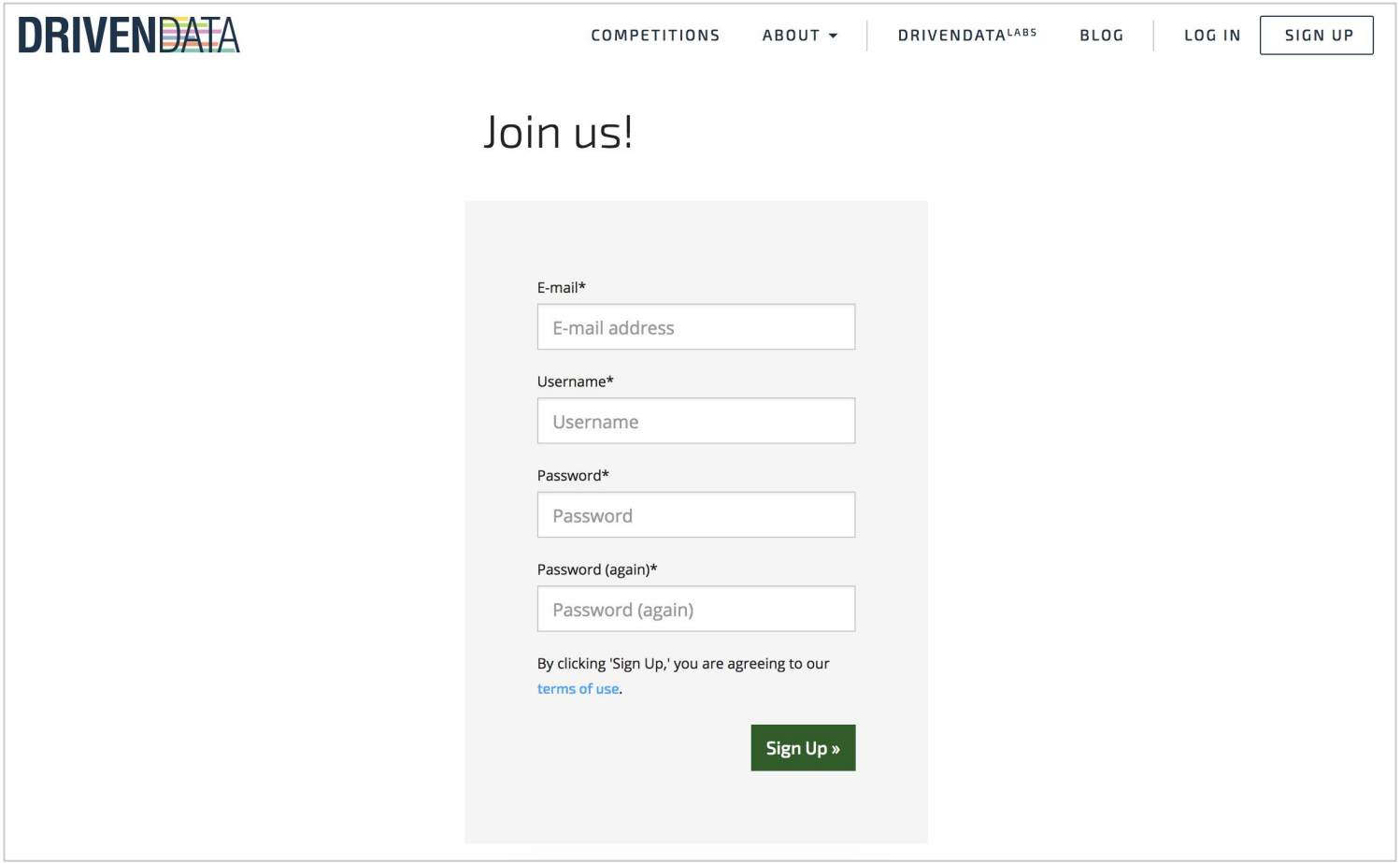
- Performs pre-processing on your data
- Uses cross validation for each model
- Uses only your **training data**

Assess your best model on your **test data**

```
# Import a classifier (a method that generates an instance of a classifier)
from sklearn.neighbors import KNeighborsClassifier

# Create a classifier
clf = KNeighborsClassifier()

# Split into test/train data
from sklearn.model_selection import train_test_split
train_features, test_features, train_outcome, test_outcome = train_test_split(data.data,
data.target, test_size=0.30)

# Grid search with preprocessing
pipe = make_pipeline(MinMaxScaler(), clf)
param_grid = {'kneighborsclassifier__n_neighbors': [1, 3, 5, 10]}

# Pass your pipeline to a grid search, specifying a set of neighbors to assess
grid = GridSearchCV(pipe, param_grid)
grid.fit(train_features, train_outcome)
grid.score(test_features, test_outcome) # Will use the best model in the grid
```

Machine learning steps (in code)

Driven Data



Sign up for an account with id **net-id-UW** (i.e., mikefree-UW) as your username (link)

Submission format

This competitions uses log loss as its evaluation metric, so the predictions you submit are the probability that a donor made a donation in March 2007.

The submission format is a csv with the following columns:

	Made Donation in March 2007
659	0.5
276	0.5
263	0.5
303	0.5
83	0.5

To be explicit, you need to submit a file like the following with predictions for every ID in the Test Set we provide:

```
,Made Donation in March 2007
659,0.5
276,0.5
263,0.5
303,0.5
...
```

Submissions

BEST SCORE	CURRENT RANK	# COMPETITORS	SUBS. TODAY
0.4472	561	3885	0 / 3

EVALUATION METRIC

$$\text{Log loss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

The metric used for this competition is logarithmic loss. $\hat{y}$ is the probability that $y = 1$. Logarithmic loss provides a steep penalty for predictions that are both confident and wrong. The goal is to minimize the log loss.

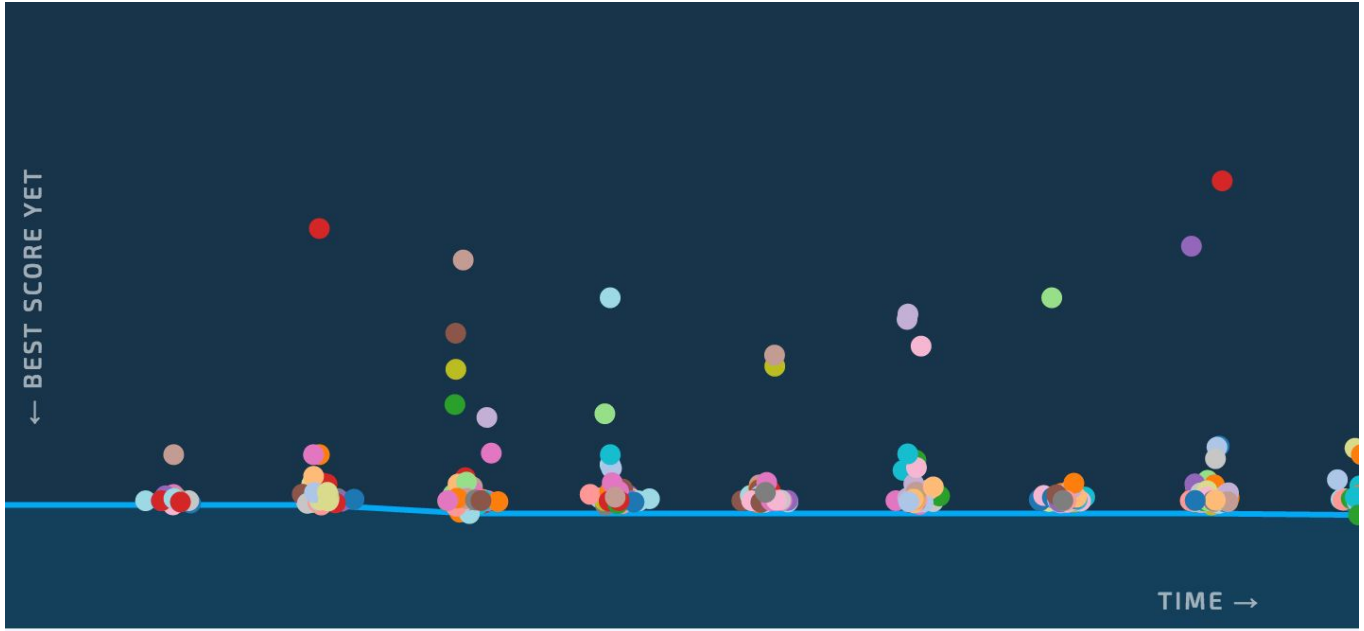
SUBMISSIONS

Score	Submitted by	Timestamp
!	mikefree	Feb. 26, 2017, 11 p.m.
!	mikefree	Feb. 26, 2017, 11:01 p.m.
!	mikefree	Feb. 26, 2017, 11:01 p.m.
9.0172	mikefree	Feb. 26, 2017, 11:03 p.m.
0.4472	mikefree	Feb. 26, 2017, 11:12 p.m.

Make new submission

Submissions are made via uploading test data

Applied Example



Driven Data Blood Donation Predictions (link)

Data/Evaluation

Here's what the first few rows of the training set look like:

	Months since Last Donation	Number of Donations	Total Volume Donated (c.c.)	Months since First Donation	Made Donation in March 2007
619	2	50	12500	98	1
664	0	13	3250	28	1
441	1	16	4000	35	1
160	2	20	5000	45	1
358	1	24	6000	77	0

EVALUATION METRIC

Log loss $= -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$

The metric used for this competition is logarithmic loss. $\hat{y}$ is the probability that $y = 1$. Logarithmic loss provides a steep penalty for predictions that are both confident and wrong. The goal is to minimize the log loss.

Hints

Prepare your data properly

Choose an evaluation technique

Select your 3 submissions wisely

Figure out a way to tune your models (figure out max_depth, n_neighbors)

Predict **probabilities**, not **outcomes** (`clf.predict_proba(test_data)[0:,1]`)

Go!

Upcoming...

Assignment-4 due **next Tuesday**

Start chipping away at **your final projects**