Resampling and Cross Validation

Sneaky Snakes: Rachel Huang and Phil Smith

Outline

Project 4:

- Running 5-fold cross validation and LOOCV on Logistic Regression, LDA, QDA, and KNN

Project 5:

- Running 5-fold cross validation and LOOCV on Bagging, Random Forest, Boosting

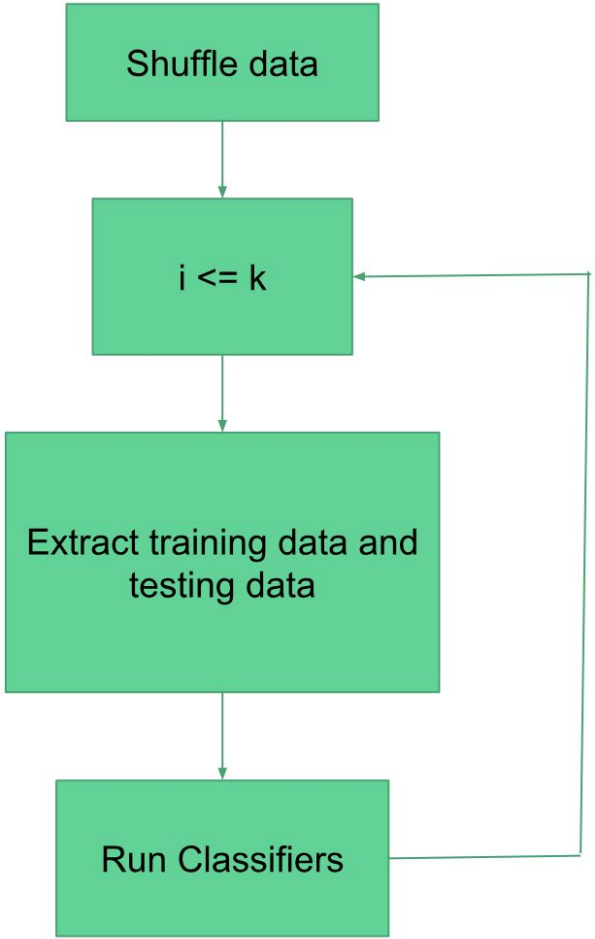
Algorithms for 5-Fold and LOOCV

5-Fold:

```
test_bif = bif[(k - 1)*200:(k * 200)]
train_bif = bif[0:((k - 1)*200)] + bif[(k * 200):]
```

LOOCV:

```
test_bif = bif[(k - 1):k]
train_bif = bif[0:(k - 1)] + bif[k:]
```



Algorithms for 5-Fold and LOOCV

```
folds = {'fold1': {'bifs': bifs[:200], 'genders': ages[:200]},
         'fold2': {'bifs': bifs[200:400], 'genders': ages[200:400]},
         'fold3': {'bifs': bifs[400:600], 'genders': ages[400:600]},
         'fold4': {'bifs': bifs[600:800], 'genders': ages[600:800]},
         'fold5': {'bifs': bifs[800:], 'genders': ages[800:]}}
```

5-Fold

For each fold:

- Make the current fold the test set
- Make the other folds the training set
- Train model
- Test model

LOOCV

For x < num_samples:

- Train = samples
- Test = samples.pop(x)
- Train model
- Test model

Gender Predictions using Various Classifiers

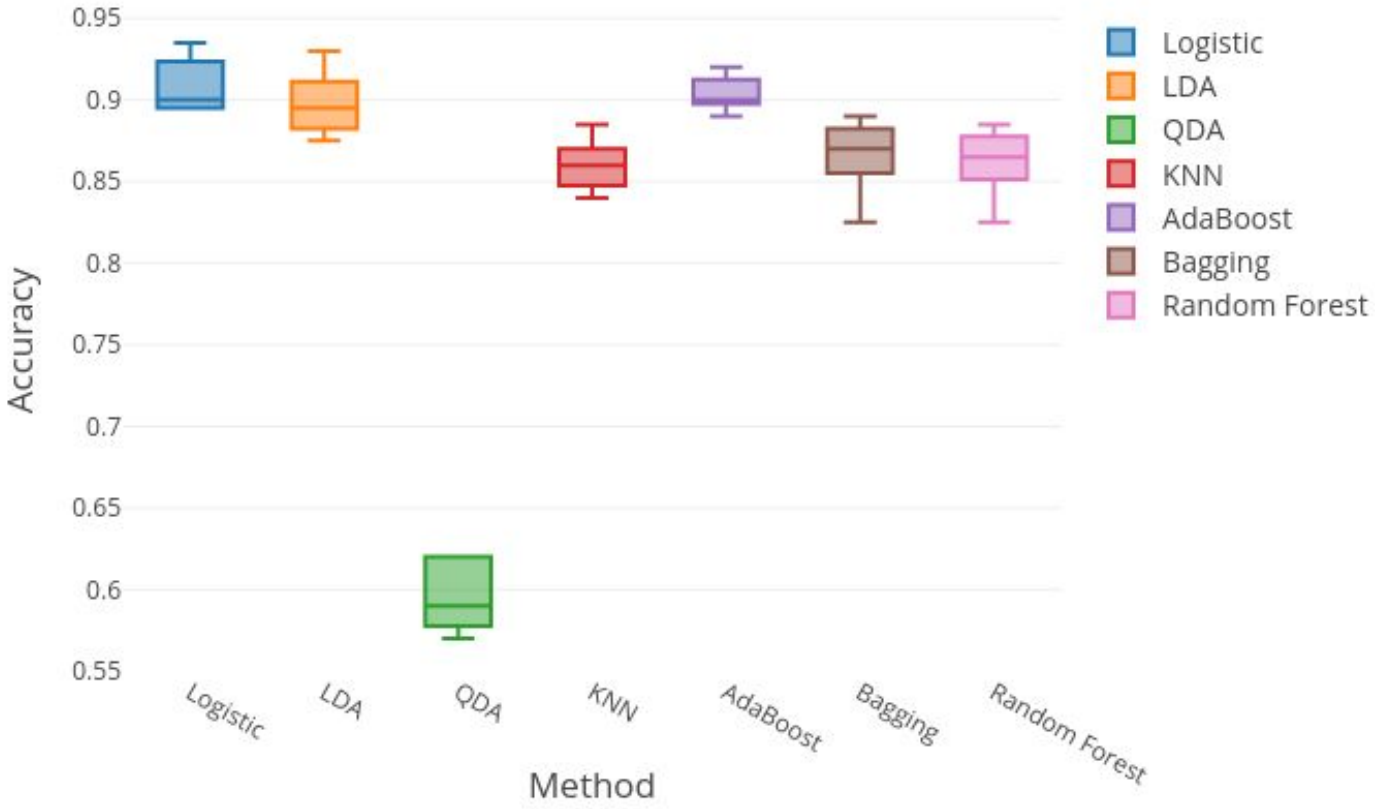
5-Fold Cross Validation

	Logistic	LDA	QDA	KNN	Boosting	Bagging	Random Forest
Mean Accuracy	**0.909**	0.898	0.596	0.860	0.904	0.866	0.862
Mean Std. Dev.	0.018	0.0211	0.023	0.017	**0.011**	0.025	0.025

Leave-one-out Cross Validation

	Logistic	LDA	QDA	KNN k=5	Boosting trees=1000	Bagging trees=50	Random Forest trees=1000
Mean Accuracy	0.908	0.898	0.574	0.861	**0.960**	0.900	0.860
Prediction Std. Dev.	0.350	0.303	0.495	**0.209**	0.313	0.303	0.347
Run Time (minutes)	30.1	≃ 9	≃ 6-8	6.3	20.1	≃ 60	≃ 30

Accuracies of Classifiers: 5-Fold Technique



5-Fold Cross Validation for Age

	Linear Regression	**AdaBoosting** Trees = 1000	**Random Forest** Trees = 1000
Mean Error (years)	8.563	13.54	10.11
Mean Std. Dev. (years)	7.0	2.37	.65
Time (min)	2.0	18.7	6.8

Questions?