# Classification pt. 2

Laura.Hangard@fysik.su.se,
Knut.Mora@fysik.su.se

# Forgotten point!

```
from sklearn.metrics import roc_curve, auc
ec = ExampleClassifier()
ec.fit(data_train,labels_train)
decisions = ec.predict_proba(data_test)[:,1]
fpr,tpr, thresholds = roc_curve(labels_test,decisions)
print "Area Under Curve is: ",auc(fpr,tpr)
```

Last time, we used ROC curves to measure performance of classifiers. If you wish to optimize a classifier, you can also use the area under the ROC curve to compute this. A perfect classifier will have AUC=1.

# Dataset

- The dataset contains 14 magnitudes for different bands, which can be subtracted to produce 91 colors

- In addition, photometric redshift and error are included

- The target will be to classify galaxies into ~ellipticals- those with 0.5<“elliptical”, and ~spirals, with “elliptical”<0.5

- Missing values are set to -9999

| Column # | Name | Description |
|---|---|---|
| 0 | counter | dummy var |
| 1 | name | SDSS galaxy ID |
| 2 | RA | Coordinates |
| 3 | DEC | same |
| 4 | elliptical | score of how elliptical |
| 5 | spiral | 1-elliptical |
| 6 | photoz | photom. redshift |
| 7 | photozErr | error on photoz |
| 8 | FUV | Magnitude, far UV |
| 9 | NUV | Magnitude, near UV |
| 10 | U | Magnitude U band |
| 11 | G | G band (optical) |
| 12 | R | R band (optical) |
| 13 | I | I band (optical) |
| 14 | Z | Z band (optical) |
| 15 | J | J band (near IR) |
| 16 | H | H band (near IR) |
| 17 | K | K band (near IR) |
| 18 | W1 | Magnitude W1-4 |
| 19 | W2 | (mid IR) |
| 20 | W3 | |
| 21 | W4 | |

# Parsing/conditioning

- To save time, a script/class called "Classify_Galaxies_Parser.py" is included
- To load the data, initialise a Galaxy_Parser("filename")
- Options may be set- see next slide

| Galaxy_Parser member | description |
|---|---|
| **data_train** | training data: one galaxy per row, one feature per column |
| **data_test** | test dataset |
| **labels_train** | array of labels for training data: 1 if spiral, 0 else |
| **labels_test** | array of labels for test data |
| **datanames** | array of name for each feature |

When initialising, some options may be set:

- precondition: if True, scale each feature vector to have 0 mean and stdev =1. Default=False

- replaceMean: if True, replace missing features of a galaxy with the average of that feature. Default = False

- ellipticity_threshold: float: if ellipticity is above this value, it will be considered a "true" elliptical galaxy. Default = 0.5

- trainfrac : float: fraction of data to use for training. The remainder will be used for test dataset Default =0.8

# Decision tree task

- Classify the galaxies using a decision tree, for different tree depths.
- Find the optimal tree depth using cross-validation
- Plot an ROC curve for the best tree depth
- explore the graph produced by the DT- for a first start, have a look at the .tree_ member of the Decision tree Classifier
- What variables are important?
- Plot two of the most important variables, as well as the cut on each
- color the points according to true spiral/ elliptical
- Experiment with preconditioning the data, or allowing the sparse data.

# Random Forest task

- Classify the galaxies using a decision tree, for different tree depths.
- Find the number of trees where the forest classifier converges
- Plot an ROC curve for the best forest
- What variables are important? (hint: see documentation for RandomForestClassifier.feature_importances_
- Plot two of the most important variables, as well as the cut on each
- color the points according to true spiral/elliptical
- Experiment with preconditioning the data, or allowing the sparse data.
- Check http://scikit-learn.org/stable/modules/ensemble.html#forest and experiment with different forest/ensemble classifiers

# Support Vector Machine task

- Classify the galaxies using a Support Vector Classifier
- What is the probability for a false and a true positive identification?
- What variables are important? (hint: what direction does the support vector point?)
- Plot two of the most important variables, color the points according to true spiral/ elliptical
- Experiment with preconditioning the data, or allowing the sparse data.
- Experiment with weighing the elliptical and spiral galaxies differently to the classifier. Can you construct an ROC curve from this, or otherwise?