MA2823: Foundations of Machine Learning — Homework 5

Due November 10, 2017.

Turn in your homework

— as a PDF file (**This means no .docx**)

— named `HW05_<LastName>_<FirstName>.pdf` (no accents)

— at `http://tinyurl.com/ma2823-2017-hw`

Question 1

What is the cross-validated **mean absolute error** of a linear regression on the wine data from Lab 4? (You can use either scikit-learn's or your own implementation.) Do you think this is good performance?

Solution:

```
from sklearn import linear_model
regr = linear_model.LinearRegression()
pred = cross_validate_regr(X_regr, y_regr, regr, folds_regr)

print("MAE: %.3f" % metrics.mean_absolute_error(y_regr,pred))
```

The mean absolute error if 0.586. Given that scores are integers on a scale from 3 to 9, it is not bad – in average, the predictor is closer to the true score than the next possible score.

Question 2

What is the cross-validated **recall** of a logistic regression on the breast cancer data from Lab 4? (You can use either scikit-learn's or your own implementation.) Do you think this is good performance?

Solution:

```
from sklearn import linear_model
from sklearn import metrics

clf = linear_model.LogisticRegression(C=1e6)
ypred_logreg = cross_validate_clf(X_clf, y_clf, clf, folds_clf)

ypred_logreg_pos = np.where(ypred_logreg > 0.5, 1, 0)

print("recall: %.3f" % metrics.recall_score(y_clf, ypred_logreg_pos))
```

Recall is close to 93%. This is good performance, although of course we have used a large number of genes to obtain this result.