Week 4 Assignment

Rubric Details

Maximum Score	50 points

i(a)
5 possible points (10%)

. Use crossvalidation to choose polynomial order q and regularisation weight C. Possible options: grid both values and choose jointly, pick q for a given C then pick C (but double check if final C is v different from initial guess), pick C for a given q then pick q. Give plots of performance (e.g. F1 score, AUC) vs q and C, incl error bars. Plot predictions and training data together to visualise performance of trained model with selected q and C. Plot should be clear so that training data and predictions are easily distinguished and +1/-1 points can be identified, should include legend identifying points (or accompanying text doing this) and legible labels/axes ticks.	**0 – 5**

i(b)
5 possible points (10%)

. As for (a) but select k for kNN classifier.	**0 – 5**

i(c)
5 possible points (10%)

.

Confusion matrices for logistic regression, kNN and baseline classifiers. Explain how these are calculated (don't just quote mystery numbers generated by using sklearn as a black box). Its ok to calc confusion matrix by splitting data into training/test sets and using test set, also fine if in parts (a) and (b) keep a hold out test set aside and then use that here in part (c), even ok to use full training data, but should mention which choice was use to calc confusion matrix. Choice of baseline classifier should be appropriate.

0 – 5

---

i(d)

5 possible points (10%)

.

Plot ROC curves for logistic regression, kNN and baseline classifiers. ROC curves for logistic regression, kNN should contain enough points to have reasonable level of detail (just one or two points is not enough), ROC curve for baseline will just be a single point. Plots should be clear, curves labelled and all text legible.

0 – 5

---

i(e)

5 possible points (10%)

.

Discuss/compare classifier performance using confusion matrices and ROC curves. ROC curves in particular are cv helpful. One dataset will be hard to predict and one will be easier, and should spot this difference in parts (i) and (ii). For hard dataset ROC curve will be close to the 45 degree line, for easier dataset it will be close to ideal (0,1) point - should observe this and comment upon this. Logistic regression and kNN classifier performance is likely much the same, but in any case their relative performance should be compared.

0 – 5

---

ii(a)

5 possible points (10%)

.

as for i(a)

0 – 5

---

ii(b)
5 possible points (10%)

.
as for i(b)

0 – 5

ii(c)
5 possible points (10%)

.
as for i(c)

0 – 5

ii(d)
5 possible points (10%)

.
as for i(d)

0 – 5

ii(e)
5 possible points (10%)

.
as for i(e)

0 – 5