You are to implement logistic regression for predicting that email’s are spam. A link to the data is provided on the class web site.

Implement the following form of cross validation:

You start with a set of models typically produced by setting a parameter. Examples of how to choose an interesting set of models are given below. Do cross validation according to the following scheme:

for i=1 to 10
  permute data
  split into 3/4 training and 1/4 testing set

Now do 5 fold cross validation to determine the best model parameters:
- partition training set into 5 parts
- for each of the 5 holdouts
  * train all models on the 4/5 part
  * record average logistic loss on 1/5 part
- The best model is chosen as the one with the best average over the 5 holdouts

Evaluate the best model by computing the average logistic loss on the 1/4 test set

Report the average performance of the best model on the 10 runs as well as the standard deviation

Which model (parameters) to optimize:

- Do $P$ passes over your data.
  - Always do a gradient descent step for the current single example.
  - At pass $i$ use learning rate $\eta_0 \alpha^{i-1}$.
    Good choices are $P = 100, \eta_0 = .2, \alpha = .95$.

For your model parameter you can optimize $P, \eta_0$, or $\alpha$.

- Versions of Shrink/Stretch. The models would be various choices for the interval $[a, b]$. Train each model with $P = 100, \eta_0 = .2, \alpha = .95$.
or some other reasonable choice. Or train it until the total gradient is less than $10^{-4}$.

- Regularized versions of logistic regression where the models are different regularization parameters. Train all models to the same precision.

Write a short report summarizing what you did including any interesting visualization.

More helpful suggestions hints:

- Before you implement cross validation, test your models on a simple split into 3/4 test and 1/4 training set.

- If your algorithms are too expensive to experiment with, then work on a subset of the data set.

- Look at the talk presented in Lecture 10. Try to reproduce the same type of experimental results.