

## Notes for Homework 1:

The homework asks you to:

Do a paired-t test to determine the percent confidence that the mean accuracy of the perceptron trained on the original data is less than (or greater than) the mean accuracy of the perceptron trained on the PCA-transformed data set. Use the algorithm described in class, with ten-fold cross-validation.

Here is some clarification on what this means. What you need to do is the following;

1. Take the training set  $D_0$  (“optdigits.tra”), and divide it into 10 disjoint subsets with approximately equal sizes:  $T_1, T_2, \dots, T_{10}$ .
2. For  $i = 1$  to 10:
  - Let  $S_i = \{D_0 - T_i\}$ .
  - Run PCA on  $S_i$  to obtain two new training sets:  $PCA(S_i)$  (with 64 features) and  $PCA'(S_i)$  (using only the first four principal components as features).
  - Use  $S_i$  as the training set to train perceptron  $h_i^{\text{raw}}$ .
  - Use  $PCA(S_i)$  as the training set to train perceptron  $h_i^{\text{PCA}}$ .
  - Use  $PCA'(S_i)$  as the training set to train perceptron  $h_i^{\text{PCA-reduced}}$ .
  - Use the PCA basis vectors you got in step 2 to create  $PCA(T_i)$  from  $T_i$ .
  - Calculate the fraction  $err_i(h_i^{\text{raw}})$  of errors made by perceptron  $h_i^{\text{raw}}$  on  $T_i$ .

- Calculate the fraction of errors  $err_i(h_i^{\text{PCA}})$  made by perceptron  $h_i^{\text{PCA}}$  on  $\text{PCA}(T_i)$ .
- Calculate the fraction of errors  $err_i(h_i^{\text{PCA-reduced}})$  made by perceptron  $h_i^{\text{PCA-reduced}}$  on  $\text{PCA}(T_i)$ .
- Calculate

$$\delta_i = err_i(h_i^{\text{PCA}}) - err_i(h_i^{\text{raw}})$$

and

$$\delta'_i = err_i(h_i^{\text{PCA-reduced}}) - err_i(h_i^{\text{raw}})$$

3. Calculate

$$\bar{\delta} = \sum_1^{10} \delta_i$$

and

$$\bar{\delta}' = \sum_1^{10} \delta'_i$$

4. We want to determine the probability that the true difference,  $\delta$ , is greater than zero, given our observation for  $\bar{\delta}$ . We also want to determine the probability that the true difference  $\delta'$  is greater than zero, given our observation for  $\bar{\delta}'$ .

To figure out the probability that  $\delta > 0$ , we rewrite it as

$$\bar{\delta} < \delta + \bar{\delta}.$$

We know for a normal distribution, we have, with confidence N%:

$$\bar{\delta} < \delta + t_{N,k-1} s_{\bar{\delta}}$$

Here I am using  $s$  as a shorthand to represent the standard deviation of the sample mean:

$$s_{\bar{\delta}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^k (\delta_i - \bar{\delta})^2}.$$

So we can set

$$\bar{\delta} = t_{N,k-1} s_{\bar{\delta}},$$

and solve for  $t_{N,k-1}$ .

We then look up the two-sided confidence value associated with this  $t_{N,k-1}$  and then turn it into a one-sided confidence interval to give us the percent confidence that  $\delta > 0$  given our observation of  $\bar{\delta}$ . See section 5.5.1 of the reading for a discussion of this. The only difference here is that we're dealing with the distribution of *means* of errors instead of simply the distribution of errors, so we use  $t_{n,k-1}$  instead of  $Z_N$ , and  $s$  instead of  $\sigma$ .