# CMSC 254: Problem Set 2

Rachel Hwang

2/7/2014

Discussed with: Andrew Ding.

1. **Problem 1**

   (a) Using the properties $\vec{v} \cdot \vec{v} = \|\vec{v}\|^2$ and $\vec{v} \cdot (\vec{u} \cdot \vec{w}) = (\vec{v} \cdot \vec{u}) + (\vec{v} \cdot \vec{w})$ and$c_1 u \cdot c_2 w = c_1 c_2 (u \cdot w)$, we know that

$$\begin{aligned}
\|P_2(\vec{x_i})\|^2 =& \|(x \cdot w_1)w_1 + (x \cdot w_2)w_2\|^2 \\
=& [(x \cdot w_1)w_1 + (x \cdot w_2)w_2] \cdot [(x \cdot w_1)w_1 + (x \cdot w_2)w_2] \\
=& [(x \cdot w_1)w_1 + (x \cdot w_2)w_2] \cdot ((x \cdot w_1)w_1) \\
& + [(x \cdot w_1)w_1 + (x \cdot w_2)w_2] \cdot ((x \cdot w_2)w_2) \\
\\
=& [(x \cdot w_1)w_1 \cdot (x \cdot w_1)w_1] \\
& + [(x \cdot w_2)w_2 \cdot (x \cdot w_2)w_2] \\
& + 2[(x \cdot w_1)w_1 \cdot (x \cdot w_2)w_2] \\
\\
=& \|(x \cdot w_1)w_1\|^2 + \|(x \cdot w_2)w_2\|^2 \\
& + 2[(x \cdot w_1)w_1 \cdot (x \cdot w_2)w_2] \\
\\
=& \|(x \cdot w_1)w_1\|^2 + \|(x \cdot w_2)w_2\|^2 \\
& + 2[(x \cdot w_1)(x \cdot w_2)](w_1 \cdot w_2) \\
\\
=& \|(x \cdot w_1)w_1\|^2 + \|(x \cdot w_1)w_1\|^2 \\
& + 2[(x \cdot w_1) \cdot (x \cdot w_2)](0) \\
\\
=& (x \cdot w_1)^2 + (x \cdot w_2)^2
\end{aligned}$$

   In the penultimate step, we know that $w_1 \cdot w_2 = 0$ since $w_1$ and $w_2$ are orthogonal. We may thus conclude that

$$\phi(\vec{w_1}, \vec{w_2}) = \frac{1}{m} \sum_{i=1}^{m} \|P_2(\vec{x_i})\|^2 = \frac{1}{m} \sum_{i=1}^{m} [(\vec{x_i} \cdot \vec{w_1})^2 + (\vec{x_i} \cdot \vec{w_2})^2]$$

(b) Since the eigen vectors form a basis, we know that $w_2 = \sum_{i=1}^{d} \alpha_i v_i$. We want to show that $\alpha_1 = 0$. We know that $w_1$ is orthogonal to $w_2$, or $w_1 \cdot w_2 = 0$.

$$\begin{aligned} w_2 =& \sum_{i=1}^{d} \alpha_i v_i \\ w_2 \cdot v_1 =& (\sum_{i=1}^{d} \alpha_i v_i) \cdot v_1 \\ =& \sum_{i=1}^{d} (\alpha_i v_i) \cdot v_1 \\ =& \sum_{i=1}^{d} \alpha_i (v_i \cdot v_1) \\ 0 =& \alpha_1 (v_1 \cdot v_1) + \sum_{i=2}^{d} \alpha_i (v_i \cdot v_1) \end{aligned}$$

Since any set of eigen vectors are mutually orthogonal, we know that for all $i > 1$, $(v_i \cdot v_1) = 0$. However, $(v_1 \cdot v_1) = 1$, therefore

$$\alpha_1 = \alpha_1 (v_1 \cdot v_1) + 0 = 0$$

(c) Beginning with the equivalence shown in 1(a),

$$\phi(\vec{w_1}, \vec{w_2}) = (\frac{1}{m} \sum_{i=1}^{m} (\vec{x_i} \cdot \vec{w_1})^2) + (\frac{1}{m} \sum_{i=1}^{m} (\vec{x_i} \cdot \vec{w_2})^2)$$

Using bilinearity, and given that the definition of the empirical covariance matrix $S = \frac{1}{m} \sum_{i=1}^{m} \vec{x_i} \vec{x_i}^\top$, we can show that

$$\begin{aligned} (\vec{x} \cdot \vec{w_2})^2 =& (\vec{x} \cdot \vec{w_2})(\vec{x} \cdot \vec{w_2}) \\ =& \vec{w_2}^\top \vec{x} \vec{x}^\top \vec{w_2} \end{aligned}$$

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^{m} (\vec{x_i} \cdot \vec{w_2})^2 =& \frac{1}{m} \sum_{i=1}^{m} \vec{w_2}^\top \vec{x} \vec{x}^\top \vec{w_2} \\ =& \vec{w_2}^\top (\frac{1}{m} \sum_{i=1}^{m} \vec{x} \vec{x}^\top) \vec{w_2} \\ =& \vec{w_2}^\top S \vec{w_2} \\ =& (\sum_{i=2}^{d} \alpha_i \vec{v_i})^\top S (\sum_{j=2}^{d} \alpha_j \vec{v_j}) \end{aligned}$$

In the penultimate step, we use the expression for $w_2$ from 1(b). We can thus conclude

$$\phi(\vec{w_1}, \vec{w_2}) = (\frac{1}{m} \sum_{i=1}^{m} (\vec{x_i} \cdot \vec{w_1})^2) + (\sum_{i=2}^{d} \alpha_i \vec{v_i})^\top S (\sum_{j=2}^{d} \alpha_j \vec{v_j})$$

(d) First, because transposition respects addition, $(A+B)^\top = A^\top + B^\top$, we know that

$$(\sum_{i=2}^{d} \alpha_i \vec{v_i})^\top = \sum_{i=2}^{d} \alpha_i (\vec{v_i}^\top)$$

Given that the spectral decompostion of $S = \sum_{i=1}^{d} \lambda_i \vec{v_i} \vec{v_i}^\top$ and using the steps shown in 1(c), we can show that

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}(\vec{x_i}\cdot\vec{w_2})^2 =& (\sum_{i=2}^{d}\alpha_i\vec{v_i}^\top)S(\sum_{j=2}^{d}\alpha_j\vec{v_j}) \\
=& (\sum_{i=2}^{d}\alpha_i\vec{v_i}^\top)(\sum_{j=1}^{d}\lambda_j\vec{v_j}\vec{v_j}^\top)(\sum_{k=2}^{d}\alpha_k\vec{v_k}) \\
=& \sum_{i=2}^{d}\sum_{j=1}^{d}\sum_{k=2}^{d}\alpha_i\vec{v_i}^\top\lambda_j\vec{v_j}\vec{v_j}^\top\alpha_k\vec{v_k} \\
=& \sum_{i=2}^{d}\sum_{j=1}^{d}\sum_{k=2}^{d}\alpha_i\lambda_j\alpha_k\vec{v_i}^\top\vec{v_j}\vec{v_j}^\top\vec{v_k} \\
=& \sum_{i=2}^{d}\sum_{j=1}^{d}\sum_{k=2}^{d}\alpha_i\lambda_j\alpha_k(\vec{v_i}\cdot\vec{v_j})(\vec{v_j}\cdot\vec{v_k})
\end{aligned}$$

However, because all the eigen vectors are orthogonal, we know that $\vec{v_x}\cdot\vec{v_y} = 0$ whenever $x \neq y$ and 1 otherwise. Thus, all the terms in the above expression are equal to zero unless $i = j = k$. this allows us to reduce to sum as follows

$$\begin{aligned}
\sum_{i=2}^{d}\sum_{j=1}^{d}\sum_{k=2}^{d}\alpha_i\lambda_j\alpha_k(\vec{v_i}\cdot\vec{v_j})(\vec{v_j}\cdot\vec{v_k}) &= \sum_{i=2}^{d}\alpha_i\lambda_i\alpha_i \\
&= \sum_{i=2}^{d}\lambda_i\alpha_i^2
\end{aligned}$$

We can thus conclude that

$$\phi(\vec{w_1},\vec{w_2}) = (\frac{1}{m}\sum_{i=1}^{m}(\vec{x_i}\cdot\vec{w_1})^2) + \sum_{i=2}^{d}\lambda_i\alpha_i^2$$

(e) We are trying to maximize the value of $\sum_{i=2}^{d}\lambda_i\alpha_i^2$. Because we are given that $\lambda_2$ is the largest lambda term, this means we need to maximize $\alpha_2$. We also know that $\sum_{i=1}^{d}\alpha_i^2 = 1$, therefore the largest any $\alpha$ term can be is 1. In this case, the optimal values of $\alpha$ would therefore be $\alpha_2 = 1$ and $\alpha_i = 0$ for all $i \neq 2$. By choosing $\alpha_2 = 1$, we are choosing $v_2$ as the principle component, $w_2 = v_2$.

2. **Problem 2** First consider that the cost function for a single cluster $C$ can be defined as

$$\phi(y) = \sum_{x\in C}\sum_{j=1}^{d}(x_j - y_j)^2$$

where $x$ and $y$ are each $d$-dimensional vectors. Given that if $y$ is a (local) minimum, it must be the case that all partial derivatives of $\phi$ with respect to $y_i$ are 0 for all $1 \leq i \leq d$, we want to show that there is a unique value for each $y_i$ that will yield 0. Taking the partial derivative at dimension $j$, it must be the case that

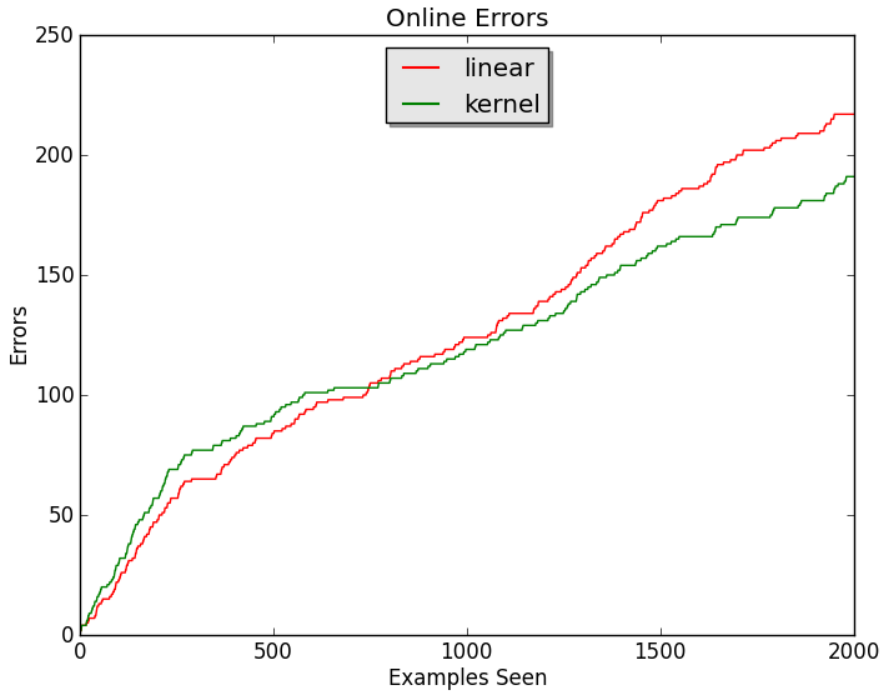
$$\begin{aligned} 0 =& \sum_{x \in C} -2(x_j - y_j) \\ =& (-2 \sum_{x \in C} x_j) + 2|C|y_j \\ y_j|C| =& \sum_{x \in C} x_j \\ y_j =& \frac{1}{|C|} \sum_{x \in C} x_j \end{aligned}$$

This shows that there since there is at most one value of each $y_j$ that satisfies the property, there is only one possible local minimum per cluster. We may extend this argument to all clusters because each cluster cost is independent.

Using the same logic, we make use the cost function defined above to find the minimum over all clusters. The minimized cost over all clusters, must be the sum of the minimal cost of each cluster. Since we have shown that there is at most one solution to the cost function per cluster, it must be the case that there is at most one (local) minimum.

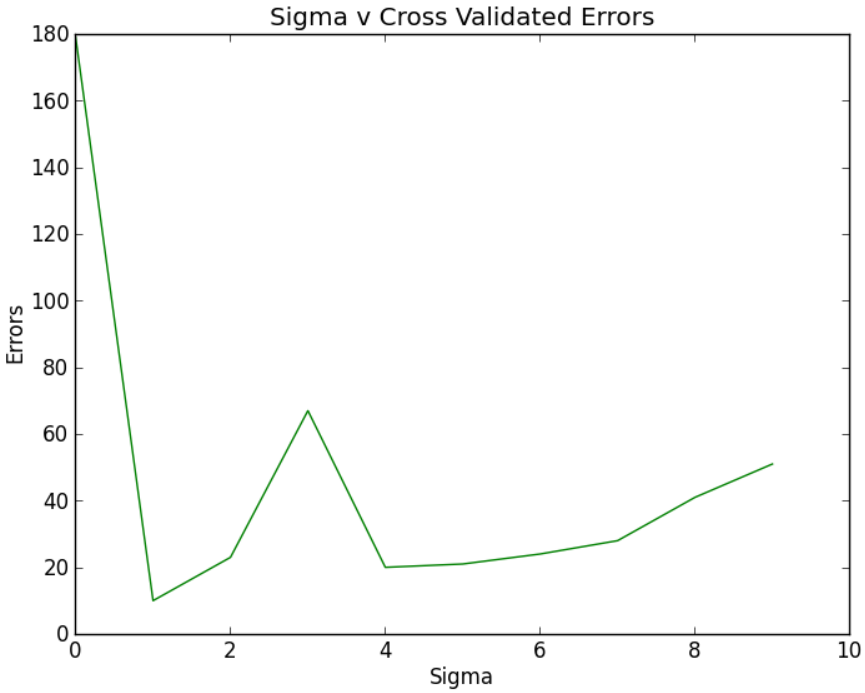
3. **Problem 3**

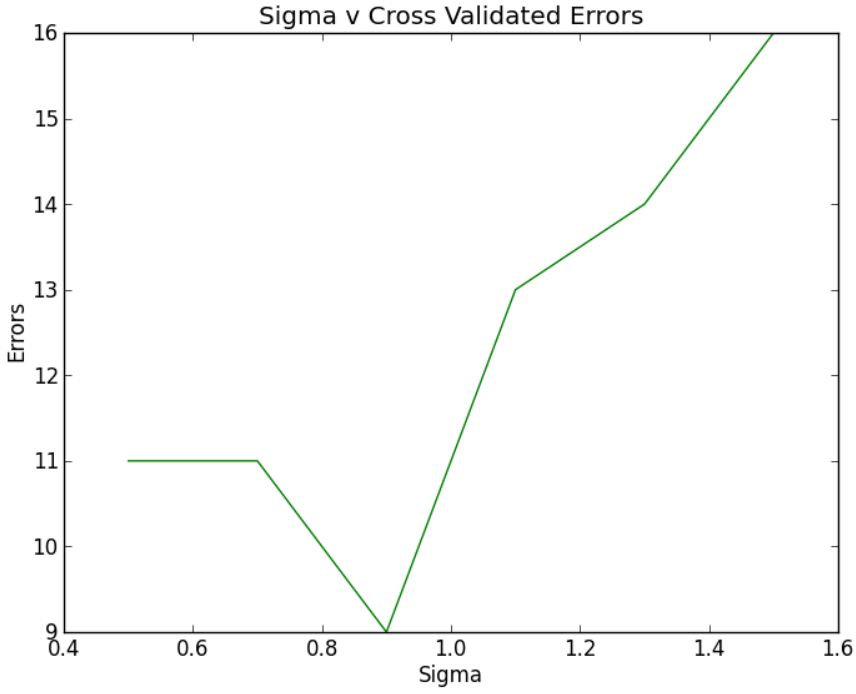
The linear and kernel perceptrons are implemented in "linear.py" and "kernel.py" respectively. Shown below are the number of errors made by both the linear and kernel programs in online mode, as a function of examples seen. The kernel program was run with a sigma value of 1. We can observe that in online mode with these sigma, the kernel and linear performances are comparable. After 2000 examples, the linear has a total accuracy rate of 89.15% and the kernel has a rate of 90.45%.




As for the batch output, all predictions were made after 20 runs on the training data, for both the kernel and linear perceptrons. The kernel was run with a optimized sigma value of 0.9.

This value was selected after testing a range of sigma values and observing the resulting cross validation errors, using 80% of the "2000K" data to train and the remaining 20% of the data to test. The sigma values at intervals of 1 were as follows. After that coarse-grained search for sigma, I ran a fine-grained search using intervals of .2. A sigma value of .9 was the clear choice.


Sigma v Cross Validated Errors
Errors
Sigma
0
20
40
60
80
100
120
140
160
180
0
2
4
6
8
10


Sigma v Cross Validated Errors
Errors
Sigma
9
10
11
12
13
14
15
16
0.4
0.6
0.8
1.0
1.2
1.4
1.6