CS725: Tutorial 6

1 Detecting spam mails

One of the fundamental tasks of machine learning is to detect spam e-mails. You are given some words and a label of +1 if it is spam or -1 if it is not. Here **1** indicates the presence[1] of word and **0** the absence of word. Assume the learning rate η is $\frac{1}{2}$. Find the seperating hyperplane using perceptron training algorithm

	area	click	your	in	singles	y
a	1	1	0	1	1	+1
b	0	0	1	1	0	-1
c	0	1	1	0	0	+1
d	1	0	0	1	0	-1
e	1	0	1	0	1	+1
f	1	0	1	1	0	-1

Tricks:
(1) Diminishing learning rate η
(2) Pick "most violating point" ie pt with $y w^T \phi(x) < 0$ but largest $|y w^T \phi(x)|$

Solution: Since this is a programming exercise. I would like you to share and discuss solutions to this and evaluate others based on your personal solutions.

One possible solution
$w_{click} = 1$, $w_{in} = -1$, $w_{singles} = 1$ $\quad (w^{(0)}, b^{(0)}) = 0$

2 Computing power of perceptrons

Perceptrons can only seperate Linearly seperable data as discussed in class. Given n variables we can have 2^{2^n} boolean functions, but not all of these can be represented by a perceptron. For example when n=2 the XOR and XNOR cannot be represented by a perceptron. Given n boolean variables how many of 2^{2^n} boolean functions can be represented by a perceptron?

Solution: `http://unbc.arcabc.ca/islandora/object/unbc\%3A6871/datastream/PDF/view`

[1] `https://preview.overleaf.com/public/vgbycngdqhgc/images/a9c18fe31ba566c1dc8ecd306bd0463d880f856b.jpeg`

3 Kernel Perceptron

Recall the proof for convergence of the perceptron update algorithm. Now can this proof be extended to the kernel perceptron?

Recall that Kernelized perceptron[2] is specified as:

$$f(x) = sign\left(\sum_i \alpha_i^* y_i K(x, x_i) + b^*\right)$$

$K(x, \alpha_i) = \phi^T(x)\phi(x_i)$

The perceptron update algorithm for the Kernelized version is:

- INITIALIZE: α=zeroes()
- REPEAT: for $< x_i, y_i >$
  - If $sign\left(\sum_j \alpha_j y_j K(x_j, x_j) + b\right) \neq y_i$
  - then, $\alpha_j = \alpha_j + 1$
  - endif

$\phi = \begin{bmatrix} \Phi \\ b \end{bmatrix}$ $K(x,y) = \Phi^T(x)\Phi(y)$

Solution: Yes, in fact kernel perceptron can be derived from the perceptron update rule as follows:

Thus all convergence results follow...

$$f(x) = sign\left((w^*)^T\phi(x)\right) = sign\left(\sum_i \alpha_i^* y_i K(x, x_i) + b^*\right)$$

$sign(w^T\phi(x_i)+b)$

$w_j \sim \left[\sum_i \alpha_i y_i \phi(x_i)\right]_j$

- INITIALIZE: $w = [0, 0, \ldots, 0, 1] \Rightarrow f(x) = sign\left((w)^T\phi(x)\right) = sign\left(\sum_i \alpha_i y_i K(x, x_i) + b\right)$

  with $\alpha_i = 0$ and $b = 1$

  Note: $\phi^T(\widehat{x})\phi(x)\widehat{y} = \widehat{y}K(\widehat{x}, x) + \widehat{y}$

  $sign\left(\sum_i \alpha_i y_i \phi^T(x)\phi(x_i) + b\right)$
- REPEAT: for each $< \widehat{x}, \widehat{y} >$
  - If $\widehat{y}w^T\phi(\widehat{x}) < 0$

    $\Rightarrow f(\widehat{x}) = sign\left((w)^T\phi(\widehat{x})\right) = sign\left(\sum_i \alpha_i y_i K(\widehat{x}, x_i) + b\right) \neq \widehat{y}$
  - then, $w' = w + \Phi(\widehat{x}).\widehat{y}$

    $w_j = \left[\sum_i \alpha_i y_i \phi(x_i)\right]_j$

    → RKHS (Reproducible Kernel Hilbert space)

    Key step.

    $\Rightarrow f(x) = sign\left((w')^T\phi(x)\right) = sign\left(\sum_i (\alpha_i y_i K(x, x_i) + \underbrace{\phi^T(\widehat{x})\phi(x)\widehat{y}}_{K(\widehat{x}, x)\widehat{y}}) + b\right)$

    $= sign\left(\sum_i \alpha_i' y_i K(x, x_i) + b'\right)$ where $\alpha_i' = \alpha_i$ for all i except that $\alpha_{\widehat{x}}' = \alpha_{\widehat{x}} + 1$ and $b' = b + \widehat{y}$
  - endif

Thus, $f(x) = sign\left((w^*)^T\phi(x)\right) = sign\left(\sum_i^* \alpha_i y_i K(x, x_i)\right)$

[2] In the original tutorial problem, b was missing. Re-introducing b helps state the equivalence of kernel perceptron to regular perceptron more easily.

$$\operatorname{sign}\left(\sum_{i=1}^{n} y_i K(x,x_i)\alpha_i + \underbrace{\phi^T(\hat{x})\phi(x)\hat{y}} + b\right)$$

$$\begin{bmatrix}\tilde{\phi}(\hat{x})\\ 1\end{bmatrix}^T \begin{bmatrix}\tilde{\phi}(x)\\ 1\end{bmatrix}\hat{y}$$

$$= \tilde{\phi}^T(\hat{x})\tilde{\phi}(x)\hat{y} + \hat{y}$$

$$= \operatorname{sign}\left(\sum_{\substack{i=1\\ x_i \neq \hat{x}}}^{n} y_i K(x,x_i)\alpha_i + \hat{y}K(x,\hat{x})\hat{\alpha} + \overbrace{K(\hat{x},x)}\hat{y} + \hat{y} + b\right)$$

$$= \operatorname{sign}\left(\sum_{\substack{i=1\\ x_i \neq \hat{x}}}^{n} y_i K(x,x_i)\underbrace{\alpha_i}_{\text{No update}} + \underbrace{(\hat{\alpha}+1)}_{\text{updates}}\hat{y}K(x,\hat{x}) + \underbrace{(b+\hat{y})}_{\text{updates}}\right)$$

4 Number of iterations for convergence of perceptron update

Prove the following:

If $||\mathbf{w}^*|| = 1$ and if there exists $\theta > 0$ such that for all $i = 1, \ldots, n$, $y_i(\mathbf{w}^*)^T\phi(\mathbf{x}_i) \geq \theta$ and $||\phi(\mathbf{x}_i)||^2 \leq \Gamma^2$ then the perceptron algorithm will make atmost $\frac{\Gamma^2}{\theta^2}$ errors (that is take atmost $\frac{\Gamma^2}{\theta^2}$ iterations to converge)

Solution:

We know that $||\mathbf{w}^*||_2^2 = 1$ and $y_i\phi(\mathbf{x}_i)\mathbf{w}^* \geq \theta$ for all i. We assume that $\mathbf{w}^{(0)} = 0$

Now consider $(\mathbf{w}^*)^T\mathbf{w}^{(k)} = (\mathbf{w}^*)^T(\mathbf{w}^{(k-1)} + y_i\phi(\mathbf{x}_i)) = (\mathbf{w}^*)^T\mathbf{w}^{(k-1)} + y_i(\mathbf{w}^*)^T\phi(\mathbf{x}_i) \geq (\mathbf{w}^*)^T\mathbf{w}^{(k-1)} + \theta \geq (\mathbf{w}^*)^T\mathbf{w}^{(k-2)} + 2\theta \geq (\mathbf{w}^*)^T\mathbf{w}^{(0)} + k\theta = k\theta$

Thus,

$$(\mathbf{w}^*)^T\mathbf{w}^{(k)} \geq k\theta$$

and because

$$||\mathbf{w}^*||||\mathbf{w}^{(k)}|| = ||\mathbf{w}^{(k)}|| \geq |(\mathbf{w}^*)^T\mathbf{w}^{(k)}|$$

Cauchy Schwarz.

we must have

$$||\mathbf{w}^{(k)}|| \geq k\theta$$

Similarly,

$||\mathbf{w}^{(k)}||_2^2 = ||\mathbf{w}^{(k-1)} + y_i\phi(\mathbf{x}_i)||_2^2 = ||\mathbf{w}^{(k-1)}||_2^2 + y_i^2||\phi(\mathbf{x}_i)||_2^2 + 2y_i(\mathbf{w}^{(k-1)})^T\phi(\mathbf{x}_i) < ||\mathbf{w}^{(k-1)}||_2^2 + \Gamma^2 < ||\mathbf{w}^{(k-2)}||_2^2 + 2\Gamma^2 < ||\mathbf{w}^{(0)}||_2^2 + k\Gamma^2 = k\Gamma^2$

since $y_i^2 = 1$ and it must have been that (as per perceptron update rule) $y_i(\mathbf{w}^{(k-1)})^T\phi(\mathbf{x}_i) < 0$

Thus,

$$||\mathbf{w}^{(k)}||_2^2 < k\Gamma^2$$

and

$$||\mathbf{w}^{(k)}||_2^2 \geq k^2\theta^2$$

which implies

$$k^2\theta^2 < k\Gamma$$

that is,

$$k < \frac{\Gamma}{\theta^2}$$

which proves our claim.

`http://www.cs.columbia.edu/~mcollins/courses/6998-2012/notes/perc.converge.pdf`