CS446: Machine Learning **Fall 2014**

Problem Set 5

Nikhil Unni *Handed In: November 7, 2014*

1. SVM

   (a) 1. $\mathbf{w} = [-1, 0]^T$
          $\theta = 0$
       2. $\mathbf{w} = [-0.5, 0.25]^T$
          $\theta = 0$
       3. I found the two closest positive/negative points, $[(-1.2, 1.6), +], [(2, 0), -]$, and found the slope between them, $\frac{1.6}{-3.2} = -\frac{1}{2}$, and the midpoint, $(0.4, 0.8)$, so the line with the farthest distance between the two points (the support vectors), has a slope of 2 with a point $(0.4, 0.8)$, giving the line $y = 2x$, which gives $w = [-2, 1]^T, \theta = 0$.

          Then, I just minimized w by halving it repeatedly, until I got $w = [-0.5, 0.25]$. This w gave $y(w^T x + \theta) = 1$ for both support vectors, so I know this is the smallest value of w I can get.

   (b) 1. $I = \{1, 6\}$
       2. $\alpha = \{\frac{5}{32}, \frac{5}{32}\}$
       3. Objective function value $= \frac{5}{32}$.

   (c) C represents how much the SVM should avoid misclassifications. In general, C controls the relative importance of maximizing the margin. For $C = \infty$, we obtain our original hyperplane that we found in (a)-2. For $C = 1$, we get a larger margin, with a higher chance of misclassification. The support vectors for $C = 1$ can now be inside the margins. For $C = 0$ has an even wider margin, with even larger misclassification.

2. Kernels

(a)

Initialize α to $\vec{0}$ of length n, where n is the number of examples.
Initialize θ to 0.
While still making mistakes (terminate after long string of successes):
    For each training example (x, y):
        if $y[(\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle) + \theta] < 0$: ($\langle a, b \rangle$ representing the inner product)
            $\alpha_i \leftarrow \alpha_i + 1$ (where i is the index of the current example (x, y))
            $\theta \leftarrow \theta + y$

(b)

$$K(x,z) = \alpha K_1(x,z) + \beta K_2(x,z)$$

Since K_1 and K_2 are both valid kernel functions, they can be represented as the dot product of two feature maps, ϕ_1 and ϕ_2 such that

$$K(x,z) = \alpha\langle\phi_1(x), \phi_1(z)\rangle + \beta\langle\phi_2(x), \phi_2(z)\rangle$$

And also such that

$$\phi_1(x) = [\phi_1(x)_1, \phi_1(x)_2, \ldots, \phi_1(x)_M]$$

$$\phi_2(x) = [\phi_2(x)_1, \phi_2(x)_2, \ldots, \phi_2(x)_N]$$

Where M and N the size of the vectors $\phi_1(x)$ and $\phi_2(x)$ respectively. Then, we can represent $K(x,z)$ by using the definition of ϕ_1 and ϕ_2 and expand out the inner products

$$K(x,z) = \alpha\sum_{i=1}^{M}\phi_1(x)_i\phi_1(z)_i + \beta\sum_{j=1}^{N}\phi_2(x)_j\phi_2(z)_j$$

Then we put α and β inside the summations as follows

$$K(x,z) = \sum_{i=1}^{M}(\sqrt{\alpha}\phi_1(x)_i)(\sqrt{\alpha}\phi_1(z)_i) + \sum_{j=1}^{N}(\sqrt{\beta}\phi_2(x)_j)(\sqrt{\beta}\phi_2(z)_j)$$

Suppose we had a feature map like :

$$\phi(x) = [\sqrt{\alpha}\phi_1(x)_1, \sqrt{\alpha}\phi_1(x)_2, \ldots, \sqrt{\alpha}\phi_1(x)_M, \sqrt{\beta}\phi_2(x)_1, \ldots, \sqrt{\beta}\phi_2(x)_N]$$

of dimension $N + M$.
Then, $\langle\phi(x), \phi(z)\rangle = \alpha\phi_1(x)_1\phi_1(z)_1 + \alpha\phi_1(x)_2\phi_1(z)_2 + \ldots \alpha\phi_1(x)_M\phi_1(z)_M + \beta\phi_2(x)_1\phi_2(z)_1 + \ldots + \beta\phi_2(x)_N\phi_2(z)_N$
Or

$$\langle\phi(x), \phi(z)\rangle = \sum_{i=1}^{M}(\sqrt{\alpha}\phi_1(x)_i)(\sqrt{\alpha}\phi_1(z)_i) + \sum_{j=1}^{N}(\sqrt{\beta}\phi_2(x)_j)(\sqrt{\beta}\phi_2(z)_j)$$

$$\langle\phi(x), \phi(z)\rangle = K(x,z)$$

Because $K(x,z)$ is an inner product of our new feature map, it is a valid kernel for all valid kernels K_1 and K_2 and all positive α and β.

(c) Before proving that $K(x,z)$ is a valid kernel, I'll prove one more property of kernels.
Note: I'm going to reuse K, x and z for this proof, they're not necessarily the same K, x and z we were given in the problem, sorry for the slight abuse of notation!

1. $K(x,z) = K_1(x,z)K_2(x,z)$, for all valid kernels K_1 and K_2

$$K(x,z) = (\langle \phi_1(x), \phi_1(z)\rangle) * (\langle \phi_2(x), \phi_2(z))$$
$$K(x,z) = (\sum_{i=1}^{M} \phi_1(x)_i \phi_1(z)_i)(\sum_{j=1}^{N} \phi_2(x)_j \phi_2(z)_j)$$
$$K(x,z) = \sum_{i=1}^{M}\sum_{j=1}^{N} \phi_1(x)_i \phi_1(z)_i \phi_2(x)_j \phi_2(z)_j$$

Let $\phi(x)_{ij} = \phi_1(x)_i * \phi_2(x)_j$ and $\phi(z)_{ij} = \phi_1(z)_i * \phi_2(z)_j$. Then:

$$K(x,z) = \sum_{i=1}^{M}\sum_{j=1}^{N} \phi(x)_{ij}\phi(z)_{ij}$$
$$K(x,z) = \langle \phi(x), \phi(z)\rangle$$

Where the dimension of ϕ is $M * N$.
Because K is the inner product of the feature map of x and z, it's a valid kernel.

Now that we have that, we can continue with the original proof.

First of all, $K_1(x,z) = x^T z$ is clearly a valid kernel, where the feature map $\phi_1(x)$ is just the identity feature map, so that $K_1 = \langle \phi_1(x), \phi_1(z)\rangle$.

Next, because of my proof right above, $(x^T z)(x^T z) = (x^T z)^2$ is a valid kernel too, since it's just the product of two valid kernels. By the same logic, $(x^T z)(x^T z)^2 = (x^T z)^3$ is also a valid kernel.

Then, by the proof from part (b), we know that the linear combination (with positive coefficients) of valid kernels is a valid kernel as well.
So $1(x^T z)^3 + 400(x^T z)^2$ is valid, and then $1(x^T z)^3 + 400(x^T z)^2 + 100x^T z$ is a valid kernel as well. ■

3. Boosting

   (c) For $t = 0$ mistakes were made on the 9^{th} and 10^{th} examples, giving $\epsilon_0 = 0.2$.
   Next, $\alpha_0 = \frac{1}{2}log_2(\frac{1-\epsilon_0}{\epsilon_0}) = \frac{1}{2}log_2(4) = 1$

$$D_1 = \frac{D_0(i)}{z_t} * 2^{-\alpha_0} \text{ if } y_i = h_0(x_i)$$
$$D_1 = \frac{D_0(i)}{z_t} * 2^{\alpha_0} \text{ if } y_i \neq h_0(x_i)$$

So then, for the first 8 examples, $D_1 = \frac{\frac{1}{10}}{z_t} * 2^{-1}$, and for the last 2, $D_1 = \frac{\frac{1}{10}}{z_t} * 2^1$
This means $\frac{8}{20z_t} + \frac{2}{5z_t} = 1$, making $z_t = \frac{4}{5}$.
All together, for the first 8 examples, $D_1 = \frac{1}{16}$, and for the last two, $D_1 = \frac{1}{4}$.

i	Label	Hypothesis 1: D_0	Hypothesis 1: $x_1 \equiv [x > 5]$	Hypothesis 1: $x_2 \equiv [y > 6]$	Hypothesis 1: $h_1 \equiv [x_1]$	Hypothesis 2: D_1	Hypothesis 2: $x_1 \equiv [x > 3]$	Hypothesis 2: $x_2 \equiv [y > 8]$	Hypothesis 2: $h_2 \equiv [x_2]$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	−	1/10	−	+	−	1/16	−	+	+
2	−	1/10	−	−	−	1/16	+	−	−
3	+	1/10	+	+	+	1/16	+	−	−
4	−	1/10	−	−	−	1/16	+	−	−
5	−	1/10	−	+	−	1/16	−	+	+
6	+	1/10	+	+	+	1/16	+	−	−
7	+	1/10	+	+	+	1/16	+	+	+
8	−	1/10	−	−	−	1/16	+	−	−
9	+	1/10	−	+	−	1/4	+	+	+
10	−	1/10	+	+	+	1/4	+	−	−

Table 1: Table for Boosting results

(d) Mistakes were made on examples 1, 3, 5, and 6, giving $\epsilon_1 = \frac{4}{16} = \frac{1}{4}$.
So then $\alpha_1 = \frac{1}{2} log_2(\frac{1-0.25}{0.25}) \approx 0.79248$.
This makes the final $H(x) = sgn[1(x > 5) + 0.79248(y > 8)]$.

(e) The ϵ_0 be the original error over the distribution D_t. It's just the sum of all $D_t(i)$ over all indeces where there was a misclassification.

$\epsilon_0 = \sum_{i \in I} D_t(i)$, where I is the set of all indices where there was a misclassification.

Then, let ϵ_1 be the new error of the previous hypothesis at time t over the new distribution D_{t+1}. Because we have the same hypothesis, we'll be iterating over the same indices as before. So:

$$\epsilon_1 = \sum_{i \in I} D_{t+1}(i)$$

Going by the update formula, keeping in mind these examples are all misclassifications, such that the signs are all the same:

$$\epsilon_1 = \sum_{i \in I} \frac{D_t(i)}{z_t} * e^{\frac{1}{2} ln(\frac{1-\epsilon_0}{\epsilon_0})}$$

$$\epsilon_1 = \sum_{i \in I} \frac{D_t(i)}{z_t} (\frac{1-\epsilon_0}{\epsilon_0})^{\frac{1}{2}}$$

$$\epsilon_1 = \frac{\sum_{i \in I} D_t(i)}{z_t} (\frac{1-\epsilon_0}{\epsilon_0})^{\frac{1}{2}}$$

$$\epsilon_1 = \frac{\epsilon_0}{z_t} (\frac{1-\epsilon_0}{\epsilon_0})^{\frac{1}{2}}$$

$$\epsilon_1 = \frac{\epsilon_0}{2[\epsilon_0(1-\epsilon_0)]^{\frac{1}{2}}}(\frac{1-\epsilon_0}{\epsilon_0})^{\frac{1}{2}}$$
$$\epsilon_1 = \frac{\epsilon_0}{2 * \sqrt{(\epsilon_0)}\sqrt{(\epsilon_0)}}$$
$$\epsilon_1 = \frac{1}{2}\blacksquare$$

4. Probability

   (a) i. The expected number of children per family in town A is just 1 since they always stop having children after having their first one.
   In town B, the $P(X=1)$ is the probability of having a boy on the first go, which is 0.5, and $P(X=2)$ is the probability that they had a girl first, then a boy, which is $0.5{*}0.5$. In general, $P(X=n)=0.5^n$. So $E[X]=\sum_{i=1}^{\infty} i{*}0.5^i = 2$.
   So the expected number of children per family in town B is 2.

   ii. Let X be the random variable denoting number of boys, and Y denoting number of girls.
   The expected number of boys per family in town A is just 0.5, and the expected number of girls is 0.5 as well. This is because they only have 1 child, and it's an even shot at which gender it becomes. $P(X=1)=0.5$, $P(X=0)=0.5$, and likewise $P(Y=1)=0.5$, so the expected value for boys and girls are both $1*0.5=0.5$.

   In town B, $P(X=1)=1$, since they refuse to stop having children until a son comes. But since they stop right after, $P(X>1)=0$. This means $E[X]=1$. For girls, $P(Y=1)=0.5*0.5$ since it means the first birth was a girl, and was immediately followed by a boy. Continuing, $P(Y=2)=0.5^3$, and so on.
   So $E[Y]=1*0.5^2+2*0.5^3+\ldots=\sum_{i=1}^{\infty} i*0.5^{i+1}=1$.

   Putting it all together, the ratio in town A is $\frac{0.5}{0.5}=\mathbf{1}$, and the ratio in town B is $\frac{1}{1}=\mathbf{1}$, maintaining the existing ratio in both towns.