Machine Learning
Assignment 3.1

Nikolaj Dybdahl Rathcke (rfq695)

January 4, 2016

1 Support Vector Machines

1.1 Kernel-induced metric

We want to show that

$$\|\Phi(x) - \Phi(z)\| = \sqrt{k(x,x) - 2k(x,z) + k(z,z)} \tag{1}$$

We can exchange the right hand side with the left one and square both sides to get:

$$\begin{aligned} k(x,x) - 2k(x,z) + k(z,z) &= \|\Phi(x) - \Phi(z)\|^2 \\ &= \langle \Phi(x), \Phi(z) \rangle^2 \\ &= \langle \Phi(x), \Phi(x) \rangle - 2\langle \Phi(x), \Phi(z) \rangle + \langle \Phi(z), \Phi(z) \rangle \\ &= k(x,x) - 2k(x,z) + k(z,z) \end{aligned}$$

We can do the last step as $\langle \Phi(x), \Phi(x') \rangle = k(x,x')$, and thus we have proven equation 1.

1.2 SVM in practice

1.2.1 Data normalization

Running `svm.py` will produce the following mean and variance for the training data:

```
Calculating mean and variance
[array([  1.55960388e+02,   2.04821194e+02,   1.15058622e+02,
          5.99785714e-03,   4.28877551e-05,   3.20418367e-03,
          3.31540816e-03,   9.61295918e-03,   2.77400000e-02,
          2.62408163e-01,   1.46761224e-02,   1.66144898e-02,
          2.19880612e-02,   4.40281633e-02,   2.26390816e-02,
          2.20007041e+01,   4.94819602e-01,   7.15689765e-01,
         -5.76372753e+00,   2.14795724e-01,   2.36576287e+00,
          1.99708816e-01]),

          array([  1.96280920e+03,   9.63381571e+03,   2.09357394e+03,
          1.56323454e-05,   9.05262911e-10,   5.59068556e-06,
          5.17943912e-06,   5.03013168e-05,   2.52776890e-04,
          2.64697314e-02,   7.48602666e-05,   1.02539466e-04,
          1.76683701e-04,   6.73690070e-04,   8.86782604e-04,
          1.65097280e+01,   1.03107075e-02,   3.11595818e-03,
          1.06174931e+00,   5.74533305e-03,   1.36459832e-01,
          6.65889223e-03])]
```

The code has been tweaked slightly from assignment 1. The first array is the mean of the 22 features from the training set. The second array is the variance.
After normalizing the test data using the mean and variance from the training data, we calculate the mean and variance of the test data:

```
Mean and variance when the training set is normalized:
[array([-0.07857931, -0.15804162,  0.05562311,  0.11318387,  0.07157377,
        0.08691489,  0.11567239,  0.08701553,  0.24898214,  0.24518734,
        0.2295662 ,  0.25089051,  0.31660826,  0.22960283,  0.14905702,
       -0.05676346,  0.07356766,  0.08676698,  0.15477245,  0.31069455,
        0.08741643,  0.1685766 ]),

        array([ 0.73218508,  0.71491336,  0.79759033,  1.99040214,  1.66604029,
        2.13673681,  1.92226228,  2.13767335,  1.77195651,  1.82895633,
        1.7173149 ,  1.77783879,  2.19022855,  1.7174543 ,  2.66297002,
        1.36090146,  1.08263293,  0.95130846,  1.21651005,  1.36280271,
        1.13351689,  1.41470112])]
```

Again the first array is the mean which should be close to 0, and the second is the variance which should be close to 1.

1.2.2 Model selection using grid-search

Not implemented.

1.2.3 Inspecting the kernel expansion

As the parameter C decreases, the number of bounded support vectors will increase, and the number will decrease as C increases.
The number of free vectors will increase as C increases, and decrease as C will decrease.
This is because, as we increase C, we get a smaller margin but more correct classification. If we decrease it, the larger a margin we will have.

2 The growth function

2.1

We want to prove that

$$m_{\mathcal{H}}(N) \leq min(M, 2^N)$$

This can be rewritten into two proofs:

$$m_{\mathcal{H}}(N) \leq M \text{ and } M \leq 2^N \tag{2}$$

$$m_{\mathcal{H}}(N) \leq 2^N \text{ and } 2^N \leq M \tag{3}$$

Dichotomies are *possible distinct* ways to fill a tuple of N points with ±1's.
Now for (1), if $M \leq 2^N$ that means we have M hypotheses and we can at most have this many dichotomies, but we will not have all possible dichotomies.
For (2), when $2^N \leq M$ it means we have more hypotheses than we have ways to fill our tuple of N points. This means we will have duplicate dichotomies, but we cannot have more than 2^N different ones.
Using this, we can conclude that we can have at most $min(M, 2^N)$ dichotomies.

2.2

We want to prove that

$$m_{\mathcal{H}}(2n) \leq m_{\mathcal{H}}(n)^2 \tag{4}$$

We let

$$m_{\mathcal{H}}(n) = c \tag{5}$$

If we partition a set of $2n$ points into sets of n points, then that means we get c dichotomies in each of these sets. These two sets can, at maximum produce dichotomies corresponding to the cross product of the two sets of dichotomies. That means the LHS of equation 4 has the upper bound

$$\begin{aligned} m_{\mathcal{H}}(2n) &\leq c \cdot c \\ &= c^2 \\ &= m_{\mathcal{H}}(n)^2 \end{aligned}$$

Last step using equation 5, and thus equation 4 is proved.

2.3

We want to prove that

$$\sum_{i=0}^{d} \binom{N}{i} \leq N^d + 1 \tag{6}$$

when $d \leq N$ We begin by setting $d = 0$ and get

$$\begin{aligned} \sum_{i=0}^{0} \binom{N}{i} &\leq N^0 + 1 && \Leftrightarrow \\ 1 &\leq 1 + 1 \end{aligned}$$

which holds. Now we do our induction step, assuming equation 6 (hypothesis) holds for d we want to prove that it holds for $d+1$

$$\sum_{i=0}^{d+1} \binom{N}{i} \leq N^{d+1} + 1$$

for $d \geq 1$. We can rewrite this to

$$\begin{aligned} \sum_{i=0}^{d} \binom{N}{i} + \binom{N}{d+1} &\leq N^{d+1} + 1 \\ &= N^d + (N-1)N^d + 1 \end{aligned}$$

Using our hypothesis we can write

$$\begin{aligned} N^d + 1 + \binom{N}{d+1} &\leq N^d + (N-1)N^d + 1 && \Leftrightarrow \\ \binom{N}{d+1} &\leq (N-1)N^d \end{aligned}$$

The right hand side will always be larger since we are only looking at $d \geq 1$. Therefore equation 6 holds.

2.4

From [LFD] equation (2.9):

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d} \binom{N}{i}$$

This means we can derive the following bound

$$m_{\mathcal{H}}(N) \leq N^d + 1$$

using our result from (2.3).

2.5

Equation 2.12 in [LFD]:

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

Substituting the result in the equation yields

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4(2N)^d + 1)}{\delta}}$$

The coefficient d should be relatively small compared to N so the bound is not too loose. The smaller d compared to N, the tighter the bound.

3 VC-dimension

3.1

The VC-dimension is 3. No matter how you label the points that makes an equilateral triangle, it can be shattered. However, if you have four points, they cannot be shattered if you label the points of one diagonal one thing, and the other diagonal a second thing, they cannot be shattered. You will always have to arrange them in a quadrilateral, since if 3 points are arranged in a line it is not possible (two outer points are labeled positive). One of the circles surrounding one diagonal has to go "between" the points of the other diagonal (or exactly on if it's a square, in which case they cannot be shattered). This forces the circle surrounding the points of the other diagonal to contain all four points. If you labeled the points of the larger circle positive, it cannot be shattered, thus $d_{VC}(\mathcal{H}_+) = 3$.

3.2

The VC-dimension is 4. No matter how you label the points that makes a quadrilateral where you can make a circle of only the points that make one diagonal, it can be shattered. Since you can just use either a negative circle or positive circle depending on what the smaller diagonal is labeled. This was the only problem in (3.1), as all other labelings can be shattered by only positive circles.
However, if you have 5 points, if you use two circles, one negative and one positive, the problem is we will have some kind of triangle to just shatter 3 points. The last 2 points must be placed on different sides of the triangle (so they cannot be the only two point in a circle), since otherwise we would have another triangle in the "middle". Now label these two points the same, say positive. Now one of the points in the triangle can be labeled negative so it cannot be shattered, since at least one of these points must be included in the circle surrounding the 2 outer points. If we used two different positive circles, they would missclassify each other, therefore we must have $d_{VC}(\mathcal{H}_+) = 4$.