Machine Learning Problemset 4

Aimee Barciauskas

12 March 2016

Problem 17

Let $(x_1, y1), ..., (x_n, y_n)$ **be data in** $\mathbb{R}^d \times \{-1, 1\}$. Suppose the data are *linearly seperable...*

Since the data are linearly seperable, we can define a separating hyperplane:

$$\left\{x : f(x) = w^T x = 0\right\} where |w| = 1$$

The seperating hyperplane returns a signed distance to the plane for each x, and classification is done according to the sign:

$$f(x) : sgn\left[w^T x\right]$$

The margin of this classifier is as stated in the problem:

$$\gamma(w) = min_i \frac{y_i w^T x_i}{|w|}$$

An optimal $f(x)$ is one which maximizes the margin γ, i.e.:

$$\gamma* = max \gamma_{w,|w|=1} s.t. y_i w^T x_i \geq \gamma$$

The constratint on the norm of w, $|w| = 1$ can be removed when:

$$\frac{y_i w^T x_i}{|w|} \geq \gamma$$

and by setting $|w| = \frac{1}{\gamma}$ the maximazation optimization problem becomes a convex minimization problem of the form:

$$min_w |w| s.t. y_i w^T x_i \geq 1$$

Part 2

HELP

w is a linear combination of all the points that are at the margin distance so w is a linear combination of these 2.

$$w = \sum_i \gamma_i w$$

Problem 18

Part 1

$$y_n \in \{-1, 1\}$$
$$x_n \in \{0, 1\}$$

Data are linearly seperable by the weight vector:

$$w = (w_0, ..., w_d)$$

Being linearly seperable means, by definition, there exists such a $w^T x_i > 0$ whenever $y_i = 1$ and $w^T x_i \leq 0$ whenever $y_i = -1$

By the statement of the problem, $w^T x_i > 0$ whenever at least one $x_{i,j} = 1$. The w satisfying this is the w where w_0 is $(-1, 0)$ and all $w_1 = ... = w_d = 1$:

$$\left\{ w : -1 < w_0 < 0 and w_1 = ... = w_d = 1 \right\}$$

So $w^T x$ will be greater than 0 where at least one x_i is 1 and w_0 otherwise. The x_i's nearest the w vector are the x_i's where all elements are zero and their counterparts: those with only one dimension being equal non-zero, i.e. the x_i's satisifyiing:

$$\sum_{m=1}^{d} x_{m,i} = 1 \text{ } d \text{ is the dimension of } x$$

These two types of points define the location of the hyperplane. The margin maximizing the distance between between such x_i's is where the margin equivalent for both these points:

$$\frac{y_j w^T x_j}{|w|} = \frac{y_i w^T x_i}{|w|}$$

Where $y_j = -1$, $y_i = 1$ and $\sum_{m=1}^{d} x_{m,j} = 0$, $\sum_{m=1}^{d} x_{m,i} = 1$

To solve for the optimal $w*$, where we know all $w_1 = ... = w_d = 1$ so we are just solving for w_0:

$$w_0 = -\frac{1}{2}$$

Plugging this back into the equation for γ:

$$\gamma * (w) = \frac{1}{2\sqrt{\frac{1}{4} + d}}$$

Part 2

$$w^T x > 0 \text{ when } \sum_i x_i > \frac{d}{2}$$

$w^T x = w_0^* + k w_{1,\ldots,d}^*$ where k is the number of non-zero elements of x and $w*$ is the optimal weight vector

$$w^T x > 0 \text{ when } k \geq \frac{d+1}{2} \text{ and}$$

$$w^T x < 0 \text{ when } k < \frac{d-1}{2}$$

Say $-w_0^* = k w_{1,\ldots,d}^*$ and we can set $w_{1,\ldots,d}^* = 1$ and there are two scenarios:

1. $w_0^* + \frac{d+1}{2} > 0$ when $k \geq \frac{d+1}{2}$
2. $w_0^* + \frac{d-1}{2} < 0$ when $k < \frac{d-1}{2}$

So the possible values of w_0 are:

$$\frac{d-1}{2} < -w_0^* < \frac{d+1}{2}$$

x_i's nearest the classifying hyperplane, are those where:

$$\sum_{m=1}^{d} x_{ij} = \frac{d-1}{2} \text{ or}$$

$$\sum_{m=1}^{d} x_{ij} = \frac{d+1}{2}$$

I.e.:

$$y_i = 1, \sum_i x_i = \frac{d+1}{2}$$

$$y_j = -1, \sum_j x_j = \frac{d-1}{2}$$

$\gamma(w)$ is the same for all x_i so:

$$y_i w^T x_i = y_j w^T x_j$$

Substituting y_i, y_j with $-1, 1$:

$$w^T x_i = -w^T x_j$$

$$w_0 + \sum_{m=1}^{d} w_m x_{i,m} = -w_0 - \sum_{m=1}^{d} w_m x_{j,m}$$

$$w_0 + \frac{d+1}{2} = -w_0 - \frac{d-1}{2}$$

$$w_0 = -\frac{d}{2}$$

Plugging this into the equation for γ:

$$\gamma(w) = \frac{w_0 + \frac{d+1}{2}}{\sqrt{w_0^2 + d}}$$
$$\gamma(w) = \frac{1}{2\sqrt{\frac{d^2}{4} + d}}$$

Problem 19

Part 1

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

where:

$$\phi(x)_n = \frac{1}{\sqrt{n!}} x^n e^{\frac{-x^2}{2}}$$

$$\langle \phi(x), \phi(y) \rangle = \sum_{n=0}^{\infty} \frac{1}{\sqrt{n!}} x^n e^{\frac{-x^2}{2}} \frac{1}{\sqrt{n!}} y^n e^{\frac{-y^2}{2}}$$

Simplifying and using that $\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$:

$$= e^{xy - \frac{1}{2}(x^2+y^2)}$$

Multiply the exponent by $\frac{2}{2}$:

$$= e^{-\frac{(x-y)^2}{2}}$$

Is this the gaussian kernel?

Part 2

Generalize to

$$\mathbb{R}^d$$

$$= \sum_{i=1}^{\infty} \frac{1}{n!} (\sqrt{(x^T x)} \sqrt{y^T y})^n exp(-\frac{1}{2}(x^T x) - \frac{1}{2}(y^T y))$$

$$= exp(sqrt(x^T x y^T y) - \frac{1}{2} x^T x - \frac{1}{2} y^T y)$$

$$= exp(x^T y - \frac{1}{2} x^T x - \frac{1}{2} y^T y)$$

$$= exp(\frac{2x^T y - x^T x - y^T y}{2})$$

$$= exp(-\frac{(norm(x-y))^2}{2})$$

Problem 20

Part 1

$$K(x,y) = \left\langle \phi(x), \phi(y) \right\rangle$$
$$= \left\langle \binom{(}{\phi}_1 (x), \phi_2(x)), \binom{(}{\phi}_1 (y), \phi_2(y)) \right\rangle$$
$$= \left\langle \phi_1(x), \phi_1(y) \right\rangle + \left\langle \phi_2(x), \phi_2(y) \right\rangle$$
$$= K_1(x,y) + K_2(x,y)$$

Part 2

$$K(x,y) = \phi(x)^T \phi(y)$$
$$= \phi_1(x)^T \phi_1(y) \phi_2(x)^T \phi_2(y)$$
$$= \big(\phi_1(x)\phi_2(x)\big)^T \big(\phi_1(y)\phi_2(y)\big)$$
$$= K_1(x,y) K_2(x,y)$$

Problem 21

Part 1

There are 2^m possible substrings s, so we define the dimension to be $\mu = 2^m$

$$\phi(x)_\mu = \sum_{i=1}^{\mu} \mathbb{I}_{s_i substring of x}$$

$$\left\langle \phi(x)_\mu, \phi_\mu(y) \right\rangle = K(x,y)$$

$$= \sum_{i=1}^{2^m} \mathbb{I}_{s_i \in x} \mathbb{I}_{s_i \in y}$$
$$= \sum_{i=1}^{2^m} \mathbb{I}_{s_i \in x,y}$$
$$= \sum_{s_i \in 0,1^m} \mathbb{I}_{s_i \in x,y}$$

Part 2

Let J be the set of all possible s, the magnitude of J is the dimension of the $\mathcal{H}$ the hilbert space of this kernel function, that is the dimension of the Hilbert space of the string kernel is 2^m, i.e.:

$$\phi(x) = \mathbb{R}^n \to \mathbb{R}^\mu$$