COMPGI07: Assignment 3

Antonio Remiro Azócar, MSc Machine Learning

17 January 2017

Exercise 1

(a) Letting a be a real parameter, there is a solution to the above equations for $a \neq 1$. ($a = 1$ does not give a solution since it leads to $x + y = 1 \neq 0 = x + y$ in the above system of equations).

A unique solution, $x = 0, y = 1$, is obtained with $a = 0$. Once a is set to zero, x can only equal zero (from $x + ay = 0$). Hence, from $x + y = 1$, y can only equal one.

(b) Square matrix A is invertible since its determinant ($ad - bc = -4$) is nonzero. It is therefore not singular; a matrix is singular iff its determinant is zero. It is also not symmetric since $A = \begin{bmatrix} 1 & 3 \\ 1 & -1 \end{bmatrix} \neq \begin{bmatrix} 1 & 1 \\ 3 & -1 \end{bmatrix} = A^T$.

(c) The product between an $n \times k$ matrix and a $k \times l$ matrix is a $n \times l$ matrix. This proceeds from the usual formula for matrix products. If A is $n \times k$ and C is $k \times l$, then $B = AC$ has entries defined by,

$$b_{xy} = \sum_{z=1}^{k} a_{xz} c_{zy},$$

where b_{xy}, a_{xz} and c_{zy} are entries of A, B and C, and therefore B is a $n \times l$ matrix.

(d) The identity matrix is a diagonal matrix, whose diagonal elements are all equal to 1. This proceeds from the definition of the identity matrix, where $\mathbf{I}(X) = X$ for all vectors X.

(e) Consider the vector $x = (-3, 0, 5)$. The 1-norm of x is,

$$||x||_1 = \sum_{i=1}^{3} |x_i| = 3 + 0 + 5 = 8.$$

The 2-norm of x is,

$$||x||_2 = \Big(\sum_{i=1}^{3} |x_i|^2 \Big)^{\frac{1}{2}} = (3^2 + 0^2 + 5^2)^{\frac{1}{2}} = \sqrt{34}.$$

The ∞-norm of x is,

$$||x||_\infty = \max_{1 \leq i \leq 3} |x_i| = 5.$$

Hence, for this particular example, $||x||_\infty \leq ||x||_2 \leq ||x||_1$. We can argue that these inequalities are true for every x as follows. We have,

$$||x||_2^2 = \sum_{i=1}^{m} |x_i|^2 \leq \Big(\sum_{i=1}^{m} |x_i|^2 + 2 \sum_{i,j,i \neq j} |x_i||x_j| \Big) = ||x||_1^2,$$

$$||x||_\infty^2 = \Big(\max_{1 \leq i \leq m} |x_i| \Big)^2 = \max_{i \leq i \leq m} |x_i|^2 \leq \sum_{i=1}^{m} |x_i|^2 = ||x||_2^2,$$

since the sum of $|x_i|^2$ includes $\max_{i \leq i \leq m} |x_i|^2$. Hence we have $||x||_\infty^2 \leq ||x||_2^2 \leq ||x||_1^2$; this implies $||x||_\infty \leq ||x||_2 \leq ||x||_1$. An example of a vector where these inequalities are all tight is $x = (4, 0, 0)$, where $||x||_1 = \sum_{i=1}^{3} |x_i| = 4 + 0 + 0 = 4$, $||x||_2 = \Big(\sum_{i=1}^{3} |x_i|^2 \Big)^{\frac{1}{2}} = (4^2 + 0^2 + 0^2)^{\frac{1}{2}} = \sqrt{16} = 4$, $||x||_\infty = \max_{1 \leq i \leq 3} |x_i| = 4$.

Exercise 2

(a) Let $\mathbf{v}$ be an eigenvector of P with eigenvalue λ. Then $P(\mathbf{v}) = \lambda\mathbf{v}$. Since for a projection matrix, $P = P^2$, we have $\lambda\mathbf{v} = P\mathbf{v} = P^2\mathbf{v} = \lambda^2\mathbf{v}$. With $\mathbf{v} \neq 0$, the solutions to this are $\lambda_1 = 1$, $\lambda_2 = 0$. Hence, all of the projection's eigenvalues are 0 or 1. We will now prove that the eigenvalues of a projection are equivalent to its singular values. For a projection matrix we have $P = P^T$. Then, for the eigenvector/eigenvalue problem $P(\mathbf{v}) = \lambda\mathbf{v}$, $P^T P\mathbf{v} = \lambda^2\mathbf{v}$. Hence, λ^2 is an eigenvalue for $P^T P$, which is the square of a singular value for P. Since a projection P is positive semi-definite, $\lambda \geq 0$ and as a result $\sqrt{\lambda^2} = \lambda$. Then, the singular values are equal to the eigenvalues and are all zero or one.

(b) Let P be an orthogonal projection matrix. For a projection matrix P we have $P = P^T$. To prove that $R = I - 2P$ is orthogonal, one has to show that $RR^T = 1$. One has,

$$\begin{aligned} RR^T &= (I - 2P)(I - 2P)^T \\ &= (I - 2P)(I^T - 2P^T). \end{aligned}$$

Note that I is symmetric. P, being an orthogonal matrix, is also symmetric. Hence, $(I - 2P)(I^T - 2P^T) = (I - 2P)(I - 2P)$ and we have,

$$\begin{aligned} RR^T &= (I - 2P)(I - 2P) \\ &= I - 4P + 4P^2. \end{aligned}$$

Since $P = P^2$ for a projection matrix, $RR^T = I$ and $R = I - 2P$ is an orthogonal matrix.

Exercise 3

Consider reformulating the ridge regression problem as:

$$(w_\lambda, b_\lambda) = \text{argmin}_{w,b}\Big\{\sum_{i=1}^{m}\big(y_i - b - w^T\bar{x} - w^T(x_i - \bar{x})\big)^2 + \lambda w^T w\Big\},$$

where zero has been 'introduced' as $w^T\bar{x} - w^T\bar{x}$. One can rewrite the summation above as,

$$\sum_{i=1}^{m}\big(y_i - \hat{b} - \hat{w}^T(x_i - \bar{x})\big)^2$$

by defining 'centred' $\hat{b}$ and $\hat{w}$ as,

$$\begin{aligned} \hat{b} &= b + w^T\bar{x}, \\ \hat{w} &= w. \end{aligned}$$

Both minimisations are equivalent; if b, w minimise their respective functionals, so will $\hat{b}$, $\hat{w}$, only with a shifted intercept term. More illustratively, consider recasting the x_i to have zero mean, translating the points to the origin. In this case, the 'intercepts' b change but the 'slopes' w do not. Hence, from the above correspondences, the original ridge regression problem is equivalent to:

$$(\hat{w}_\lambda, \hat{b}_\lambda) = \text{argmin}_{w,b}\Big\{\sum_{i=1}^{m}\big(y_i - b - w^T(x_i - \bar{x})\big)^2 + \lambda w^T w\Big\}.$$

Exercise 4

If $f : \mathbb{R}^d \to [0, \infty)$ is convex:

$$f(\lambda x + (1-\lambda)y)) \leq \lambda f(x) + (1-\lambda)f(y),$$

for $x, y \in \mathbb{R}^d$. Squaring both sides,

$$f^2(\lambda x + (1-\lambda)y) \leq \lambda^2 f^2(x) + 2\lambda(1-\lambda)f(x)f(y) + (1-\lambda)^2 f^2(y).$$

Subtracting and adding the terms, $\lambda f^2(x) + (1-\lambda)f^2(y)$ (which are equivalent to $\lambda g(x) + (1-\lambda)g(y)$), gives:

$$f^2(\lambda x+(1-\lambda)y) \le \lambda^2 f^2(x)+2\lambda(1-\lambda)f(x)f(y)+(1-\lambda)^2 f^2(y)-\lambda f^2(x)-(1-\lambda)f^2(y)+\lambda f^2(x)+(1-\lambda)f^2(y).$$

Assembling together the five first terms in the RHS,

$$f^2(\lambda x+(1-\lambda)y) \le -\lambda(1-\lambda)(f(x)-f(y))^2+\lambda f^2(x)+(1-\lambda)f^2(y).$$

The first term in the RHS is never positive. Hence, the inequality can be simplified to,

$$f^2(\lambda x+(1-\lambda)y) \le \lambda f^2(x)+(1-\lambda)f^2(y).$$

The expression above is equivalent to,

$$g(\lambda x+(1-\lambda)y) \le \lambda g(x)+(1-\lambda)g(y).$$

Therefore $g(w)=f^2(w)$ is also convex for every $w \in \mathbb{R}^d$.

Exercise 5

(a) We have:

$$\text{prox}_g(w)=\text{argmin}_{u\in\mathbb{R}^d} g(u)+\frac{1}{2}||w-u||_2^2.$$

Our aim is to minimise this expression over u. To find the minimum, one solves for the root of the expression's gradient. Note we have $g(w)=||w||_1$ and so we must minimise,

$$||u||_1+\frac{1}{2}||w-u||_2^2.$$

Both terms in this expression are separable in u. Therefore, each term can be minimised individually and the expression above can be denoted as,

$$|u_i|+\frac{1}{2}(w_i-u_i)^2,$$

for all i. Consider two cases: either $u_i>0$ or $u_i<0$. If $u_i>0$, the derivative of the above expression is,

$$1-w_i+u_i=0,$$

and,

$$u_i=w_i-1. \tag{1}$$

We specified $u_i>0$; hence the expression above requires $w_i>1$. Similarly, for $u_i<0$, the derivative is,

$$-1-w_i+u_i=0,$$

and we have:

$$u_i=w_i+1, \tag{2}$$

which requires $w_i<-1$ since $u_i<0$. Additionally, in the case that $-1<w_i<1$,

$$u_i=0, \tag{3}$$

so the derivative falls between -1 and 1. Assembling (1),(2) and (3) gives the following expression for the proximity operator:

$$u_i=prox_g(w_i)=\max(|w_i|-1,0)\times\text{sign}(w_i).$$

(b) The computation changes as follows. Our aim again is to minimise expression, $\text{prox}_g(w)=\text{argmin}_{u\in\mathbb{R}^d} g(u)+\frac{1}{2}||w-u||_2^2$, over u. To find the minimum, one solves for the root of the expression's gradient. Note we now have $g(w)=||w||_1+\alpha||w||_2^2$ and so we must minimise,

$$||u||_1+\alpha||u||_2^2+\frac{1}{2}||w-u||_2^2.$$

The three terms in this expression are separable in u. Therefore, each term can be minimised individually and the expression above can be denoted as,

$$|u_i|+\alpha u_i^2+\frac{1}{2}(w_i-u_i)^2.$$

for all i. Again, we consider two cases: either $u_i > 0$ or $u_i < 0$. If $u_i > 0$, the derivative of the above expression is,

$$1 + 2\alpha u_i - w_i + u_i = 0,$$

and,

$$u_i = \frac{w_i - 1}{1 + 2\alpha}. \tag{4}$$

We specified $u_i > 0$; hence the expression above requires $w_i > 1$. Similarly, for $u_i < 0$, the derivative is,

$$-1 + 2\alpha u_i - w_i + u_i = 0,$$

and we have:

$$u_i = \frac{w_i + 1}{1 + 2\alpha}, \tag{5}$$

which requires $w_i < -1$ since $u_i < 0$. Like in exercise **(5a)**, in the case that $-1 < w_i < 1$, $u_i = 0$. Assembling (4),(5) and (3) gives the following expression for the proximity operator:

$$u_i = prox_g(w_i) = \frac{\max(|w_i| - 1, 0) \times \text{sign}(w_i)}{1 + 2\alpha}.$$

Here we can observe how the solution varies with prescribed positive parameter α. The magnitude of the solution is inversely proportional to parameter α. The α coefficient 'dampens' the solution (the denominator is always greater than one with $\alpha > 0$). As $\alpha \to \infty$, $u_i \to 0$ ($u_i = 0$ anyway if $0 \geq |w_i| - 1$). As $\alpha \to 0$, u_i tends towards its 'undamped' solution at **(a)**.