CSCI-567: Assignment #3

Due on Friday, October 16, 2015

Saket Choudhary
2170058637

Contents

Problem 1

Problem 1: (a)

Let $\sigma(a) = \frac{1}{1+e^{-a}}$ and

$$P(Y=1|X=x) = \sigma(b+w^Tx)P(Y=0|X=x) = 1-\sigma(b+w^Tx)$$

Observe that $Y=1$ when $b+w^Tx \geq 0$ and $Y=0$ when $b+w^Tx < 0$
Thus,

$$\begin{aligned}
P(Y=y|X=x) &= \sigma(b+w^T)^y(1-\sigma(b+w^Tx))^(1-y) \\
\log(P(Y=y|X=x)) &= y\log(\sigma(b+w^Tx))^y + (1-y)\log(1-\sigma(b+w^Tx)) \\
&= y\log(\frac{\sigma(b+w^T)}{1-\sigma(b+w^Tx)}) + \log(1-\sigma(b+w^Tx)) \\
&= y(b+w^Tx) + \log(\frac{e^{-(b+w^Tx)}}{1+e^{-(b+w^Tx)}}) \\
&= y(b+w^Tx) + \log(\frac{1}{1+e^{(b+w^Tx)}}) \\
&= y(b+w^Tx) - \log(1+e^{(b+w^Tx)}) \qquad ([1.1])
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}(w) &= -\log(\prod_{i=1}^{n} P(Y=y_i|X=x_i)) \\
&= -\sum_{i=1}^{n}\log(P(Y=y_i|X=x_i)) \\
&= -\sum_{i=1}^{n}\left(y_i(b+w^Tx_i) - \log(1+e^{(b+w^Tx_i)})\right)
\end{aligned}$$

Consider $(L)(w) = y(b+w^Tx) - \log(1+e^{(b+w^Tx_i)})$

$$\begin{aligned}
\frac{\partial\mathcal{L}(w)}{\partial w} &= -(xy^T) + \frac{e^{(b+w^Tx)}x}{1+e^{(b+w^Tx)}} \\
\frac{\partial^2\mathcal{L}(w)}{\partial w^2} &= 0 + \frac{\partial}{\partial w}\left(x - \frac{x}{1+e^{(b+w^Tx)}}\right) \\
\frac{\partial^2\mathcal{L}(w)}{\partial w^2} &= \frac{x(e^{(b+w^Tx)})x^T}{(1+e^{(b+w^Tx)})^2} \geq 0 \;\forall\; x \in \mathbf{R} \\
\frac{\partial^2\mathcal{L}(w)}{\partial w^2} &= x^T\sigma(b+w^Tx)(1-\sigma(b+w^Tx))x \geq 0 \qquad (1.2)
\end{aligned}$$

From (1.2) $\frac{\partial^2\mathcal{L}(w)}{\partial w^2} \geq 0$ and hence, from the definition of convex functions, $\mathcal{L}(w)$ is indeed a convex function.

Problem 1 continued on next page...

Problem 1: (b)

When the data is perfectly linearly separable, (assume first $n/2$ of the n training points belong to class 0 and the remaining to class 1), thus our regression model should assign the first $n/2$ points to class with cent percent certainity or with probability 1 and the remaining $n/2$ to class 0 with probability 1. For this to happen, $P(Y=1|X=X_1)=1$ and $P(Y=0|X=X_0)=1$ where X_1 is the set of points belonging to class 1 and X_0 is the set of points belonging to class 0.
Clearly this scenario is possible when $||w|| \longrightarrow \infty$

Problem 1: (c)

A simple example with two points would be $(0,0)$, $(1,1)$. Intuitively the step function's step branches (the horizontals of a sigmooid function) will be located at infinity. Also the line separating the points (0,0) and (1,1) can be anywhere in between 0 and 1, thus there will be multiple solutions.

Problem 1: (d)

$$\begin{aligned}
\mathcal{L}(w) &= \sum_{j=1}^{n}\big(-y_j(b+w^Tx_j)+\log(1+e^{(b+w^Tx_j)})\big)+\lambda||w||_2^2\\
\frac{\partial(\mathcal{L})(w)}{\partial w_i} &= \sum_{j=1}^{n}\big(-y_j(x_{ji})+\frac{x_{ji}e^{(b+w^Tx_j)x_{ij}}}{1+e^{(b+w^Tx_j)}}\big)+2\lambda w_i=0\\
\frac{\partial^2(\mathcal{L})(w)}{\partial w_i^2} &= \sum_{j=1}^{n}\big(\frac{x_{ji}^2e^{(b+w^Tx_j)x_{ij}}}{(1+e^{(b+w^Tx_j)})^2}\big)+2\lambda>0
\end{aligned}$$

where the last inequality holds since $\lambda > 0$ Consider $\boldsymbol{f(w_i) = \sum_{j=1}^{n}\big(-y_j(x_{ji})+\frac{x_{ji}e^{(b+w^Tx_j)x_{ij}}}{1+e^{(b+w^Tx_j)}}\big)+2\lambda w_i=0}$
And u, v are the two solutions of $f(w_i)=0$, i.e. $f(u)=f(v)=0$ (Without loss of generality, assume $u<v$)
By Rolle's theorem, If $f(u)=f(v)=0$ then there exists a point in $[u,v]$ say c such that $f'(c)=0$ for $c \in [u,v]$
But, $f'(w_i)=\sum_{j=1}^{n}\big(\frac{x_{ji}^2e^{(b+w^Tx_j)x_{ij}}}{(1+e^{(b+w^Tx_j)})^2}\big)+2\lambda>0$ and hence there exists no such c.
and hence the function is convex, thus the solution to the partial differential $\frac{\partial(\mathcal{L})(w)}{\partial w_i}$ is unique.

Problem 2

Problem 2

Problem 2: (a)

Consider $||w||_0 = \#i : w_i \neq 0$ for a 1D case. Where, $x_1 = (0)$ and $x_2 = (\epsilon)$ where $0 < \epsilon << 1$
$f(w) = \sum_i I\{w_i \neq 0\}$
Since we are in 1D:
$\boldsymbol{f(w)} = \begin{cases} \mathbf{0} & \text{if w=0} \\ \mathbf{1} & \text{otherwise} \end{cases}$
Thus,

$$\begin{aligned} f(0) &= 0 \\ f(\epsilon) &= 1 \\ f(\lambda \times 0 + (1-\lambda) \times \epsilon) &= 1 \forall 0 < \lambda < 1 \quad (2a.1) \\ \lambda f(0) + (1-\lambda) f(\epsilon) &= 1 - \lambda 0 < 1 - \lambda \qquad < 1 \quad (2a.2) \end{aligned}$$

From $(2a.1), (2a.2)$ we see:
$f(\lambda \times 0 + (1-\lambda) \times \epsilon) > \lambda f(0) + (1-\lambda) f(\epsilon)$
Thus, $||w||_0$ is not a convecx function!

Problem 2: (b)

$||w||_1 = \sum_i |w_i|$
Consider two vectors u, v(same dimension say in $\mathbf{R^D}$)
Assume: $0\lambda < 1$

$$\begin{aligned} ||\lambda u + (1-\lambda)v|| &= \sum_{i=1}^{D} |\lambda u_i + (1-\lambda) v_i| \\ &\leq \sum_{i=1}^{D} \big(|\lambda u_i| + |(1-\lambda)v|\big) \; (since \; |a+b| \leq |a| + |b| \forall \; a, b \in R) \\ &= \sum_{i=1}^{D} |\lambda||u_i| + \sum_{i=1}^{D} |1-\lambda||v_i| \\ &= \lambda ||u||_1 + (1-\lambda)||v||_1 \text{ since}(0\lambda < 1) \quad (2a.1) \end{aligned}$$

From $(2b.1)$, we see. $||\lambda u + (1-\lambda)v||_1 \leq \lambda ||u||_1 + (1-\lambda)||v||_1$
And hence, $||w||_1$ is a convex function.

Problem 2: (c)

The idea of regularisation here is to bound the hyperparameters w_i so that they are less prone to overfitting.
One simple QP equivalent is this:
$min_w \frac{1}{2}(y_1 - w^T x_1, y_2 - w^T x_2 \cdots y_n - w^T x_n)^T I (y_1 - w^T x_1, y_2 - w^T x_2 \cdots y_n - w^T x_n)$ where I is the $n \times n$ identity matrix and the constraint is:
$||w||_1 \leq \frac{1}{\lambda}$

Problem 3

Problem 3: (a)

$min_w(\sum_i(y_i - w^T x_i)^2 + \lambda||w||_2^2)$
In more compact matrix notation, let:

$$y_{n\times 1} = (y_1\ y_2\ \cdots\ y_n)^T$$
$$X_{n\times D} = (x_1^T\ x_2^T\ \cdots\ x_n^T)^T$$

This notation, reduces the above function to:
$min_w(||y - w^T X||_2^2 + \lambda||w||_2^2)$

$$\begin{aligned}
f(w) &= min_w(||y - Xw||_2^2 + \lambda||w||_2^2)\\
&= (y - Xw)^T(y - Xw) + \lambda w^T w\\
&= (y^T - w^T X^T)(y - Xw) + \lambda w^T w\\
&= y^T y - y^T Xw - w^T X^T y + w^T X^T Xw + \lambda w^T w\\
&= y^T y - {(X^T y)}^T w - w^T X^T y + w^T X^T Xw + \lambda w^T w\\
\frac{\partial f(w)}{\partial w} &= -X^T y - X^T y + 2\lambda w + (X^T Xw + (XX^T w)) = 0\\
&= 2\lambda w + 2X^T Xw - 2X^T y = 0\\
\mathbf{w}(\lambda I_D + X^T w) &= X^T y\\
\mathbf{w} &= (X^T Xw + \lambda I_D)^{-1} X^T y
\end{aligned}$$

Problem 3: (b)

$min_w(||y - w^T\Phi||_2^2 + \lambda||w||_2^2)$ From the previous part, the solution should be of similar form:
$\mathbf{w} = (\Phi^T\Phi + \lambda I_D)^{-1}\Phi^T y$
Using the identity:
$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T(BPB^T + R)^{-1}$
Thus,
$\big((\lambda I_D + \Phi^T\Phi)^{-1}\big)\Phi^T y = \Phi^T\big(\Phi\Phi^T + \lambda I_N\big)^{-1} y$
$w^* = \Phi^T(\Phi\Phi^T + \lambda I_N)^{-1} y$

Problem 3 continued on next page. . .

Problem 3: (c)

$\hat{y} = w^{*T}\Phi(x)$
$\hat{y} = \left(\Phi^T(\Phi\Phi^T + \lambda I_N)^{-1}y\right)^T\Phi(x) = y^T\left((\Phi\Phi^T + \lambda I_N)^{-1}\right)^T\Phi^T\Phi(x)$
Now using the property, $(A^{-1})^T = (A^T)^{-1}$

$$\begin{aligned}\hat{y} &== y^T\left((\Phi\Phi^T + \lambda I_N)^{-1}\right)^T\Phi^T\Phi(x)\\ &= y^T\left((\Phi\Phi^T + \lambda I_N)^T\right)^{-1}\Phi^T\Phi(x)\\ &= y^T\left((\Phi^T\Phi + \lambda I_N)\right)^{-1}\Phi^T\Phi(x)\\ &= y^T(K + \lambda I_N)^{-1}\kappa(x)\end{aligned}$$

Where $K_{ij} = \Phi_i^T\Phi_j$ and $\kappa(x) = \phi^T\phi^T(x)$

Problem 3: (d)

Kernel ridge regression is $O(n^3)$ for n data points. Linear regression was forumlated as quadratic programing and hence is $O(n^2)$.
The extra n factor comes from the formation of kernel matrix K.

Problem 4

Given: $k_1(.,.)$ and $k_2(.,.)$ are kernel function. Thus, for any vector $y \in \mathbf{R}$, $y^TKy \geq 0$ where $K_{ij} = k(x_i, x_j)$
Mercer's theorem requires K to be positive semi-definite.

Problem 4 continued on next page. . .

Problem 4: (a)

$k_3(x,x') = a_1 k_1(x,x') + a_2 k_2(x,x')$ where$a_1, a_2 \geq 0$
Since $k_1(x,x')$ is positive definite, $\forall y \in \mathbf{R}$,

$$\begin{aligned} y^T K^{(1)} y \geq 0, \\ \text{where} \\ K^{(1)}_{ij} = k_1(x_i, x'_j) \end{aligned} \tag{4a.1}$$

Similarly,

$$\begin{aligned} y^T K^{(2)} y \geq 0, \\ \text{where} \\ K^{(2)}_{ij} = k_2(x_i, x'_j) \end{aligned} \tag{4a.2}$$

Thus, from (4a.1) and (4a.2), we get

$$\begin{aligned} y^T (K^{(1)} + K^{(2)}) y \geq 0 \ \forall y \in \mathbf{R} \implies \\ y^T K^{(3)} y \geq 0 \ \forall y \in \mathbf{R} \\ \text{where} \\ K^{(3)}_{ij} = k_3(x_i, x'_j) \end{aligned}$$

Problem 4: (b)

$k_4(x,x') = f(x)f(x')$ Let $K^{(4)}_{ij} = k_4(x_i, x_j) = f(x_i)f(x'_j)$
Since $f(x)$ is a real valued function, consider $K^{(4)}$

$$K^{(4)} = \begin{bmatrix} f(x_1)f(x'_1) & f(x_1)f(x'_2) & \cdots & f(x_1)f(x'_n) \\ \vdots & & & \\ f(x_n)f(x'_1) & f(x_n)f(x'_2) & \cdots & f(x_n)f(x'_n) \end{bmatrix}$$

$$\begin{aligned} K^{(4)} = F(\vec{x})_{n\times 1} F(\vec{x})^T_{1\times n} \\ \text{where} \\ F(x)^T_{1\times n} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots f(x_n) \end{pmatrix} \end{aligned}$$

Now consider $y^T K^{(4)} y = y^T F(x) F(x)^T y = y^T F(x) (y^T F(x))^T = ||y^T F(x)||^2_2 \geq 0$
Thus, $k_2(.,.)$ is a valid kernel function!.

Problem 4: (c)

$k_5(x, x') = g(k_1(x, x'))$ where g is a polynomial with positive coefficients.
Since g has positive coefficients, $g(x) \geq 0 \forall x \geq 0$
Now consider,

$$y^T K^{(5)} y = (y_1\ y_2 \cdots y_n) \times \begin{bmatrix} g(k_1(x_1, x_1')) & g(k_1(x_1, x_2')) & \cdots g(k_1(x_1, x_n')) \\ \vdots & & \\ g(k_1(x_n, x_1')) & g(k_1(x_n, x_2')) & \cdots g(k_1(x_n, x_n')) \end{bmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$y^T K^{(5)} y = y_1 g(k_1(x_1, x_1'))y_1 + y_2 g(k_1(x_1, x_2'))y_2 + \cdots y_n g(k_1(x_n, x_n'))y_n$$

Since $g(k_1(x_i, x_j)) \geq 0$

$$y^T K^{(5)} y \geq 0 \ \forall \ y \in \mathbf{R}$$

Thus k_5 is a kernel

Problem 4: (d)

$k_6(x, x') = k_1(x, x')k_2(x, x')$
Thus, in terms of our earlier defined matrix notation, $K^{(6)} = K^{(1)} \circ K^{(2)}$ where $\circ$ denotes element wise multiplication (also known as the Hadamard product).
Since, k_1 and k_2 are valid kernel function $\exists v_i w_j$ the eigen vectors of matrix K_1 and K_2 defines such that: $K^{(1)} = \sum_i \lambda_i v_i v_i^T$ and $K^{(2)} = \sum_j \mu_j w_j {w_j}^T$
Now,

$$\begin{aligned} K^{(6)} &= K^{(1)} \circ K^{(2)} \\ &= \sum_i \lambda_i v_i v_i^T \circ \sum_j \mu_j w_j {w_j}^T \\ &= \sum_{i,j} \lambda_i \mu_j (v_i v_i^T) \circ w_j {w_j}^T \\ &= \sum_{i,j} \lambda_i \mu_j (v_i \circ w_j)(v_j \circ w_j)^T \\ &\geq 0 \end{aligned}$$

Because $(v_i \circ w_j)(v_j \circ w_j)^T = ||v_i w_j||_2^2 \geq 0$

Problem 4: (e)

$k_7(x, x') = exp(k_1(x, x'))$
Just like subpart (c), here $g(x) = exp(x)$ (it's not a polynomial, though that does not affect the derivation we came up with in part (c)). So this is immediate from part (c).