Problem 1

(a) Ridge Regression

$$L(\vec{w}) = \sum_{i=1}^{n} (\vec{w}^T \vec{x}_i - y_i)^2 + \lambda ||\vec{w}||_2^2 = \sum_{i=1}^{n} (\vec{w}^T \vec{x}_i - y_i)^T (\vec{w}^T \vec{x}_i - y_i) + \lambda \vec{w}^T \vec{w}$$

$$= \sum_{i=1}^{n} \left(y_i^T y_i - 2\vec{w}^T \vec{x}_i{}^T y_i + \vec{w}^T \vec{x}_i^T \vec{x}_i \vec{w} \right) + \vec{w}^T \lambda I \vec{w}$$

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = \sum_{i=1}^{n} \left(-2\vec{x}_i^T y_i + 2\vec{x}_i^T x_i \vec{w} \right) + 2\lambda I \vec{w}$$

$$\vec{w}_{j+1} = \vec{w}_j - c \cdot \frac{\partial L(\vec{w}_j)}{\partial \vec{w}_j} = \vec{w}_j - c \cdot \left(\sum_{i=1}^{n} \left(-2\vec{x}_i^T y_i + 2\vec{x}_i^T x_i \vec{w}_j \right) + 2\lambda I \vec{w}_j \right)$$

(b) Lasso Regression

$$L(\vec{w}) = \sum_{i=1}^{n} (\vec{w}^T \vec{x}_i - y_i)^2 + \lambda |\vec{w}| = \sum_{i=1}^{n} (\vec{w}^T \vec{x}_i - y_i)^T (\vec{w}^T \vec{x}_i - y_i) + \lambda \sum_{\alpha=1}^{d} |w_\alpha|$$

$$= \sum_{i=1}^{n} \left(y_i^T y_i - 2\vec{w}^T \vec{x}_i{}^T y_i + \vec{w}^T \vec{x}_i^T \vec{x}_i \vec{w} \right) + \lambda \sum_{\alpha=1}^{d} |w_\alpha|$$

Because $L1$ regularizer $\sum_{\alpha=1}^{d} |w_\alpha|$ is not differentiable, the subgradient is

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = \sum_{i=1}^{n} \left(-2\vec{x}_i^T y_i + 2\vec{x}_i^T x_i \vec{w} + \lambda \, sign(\vec{w}) \right)$$

$$\vec{w}_{j+1} = \vec{w}_j - c \cdot \frac{\partial L(\vec{w}_j)}{\partial \vec{w}_j} = \vec{w}_j - c \cdot \left(\sum_{i=1}^{n} \left(-2\vec{x}_i^T y_i + 2\vec{x}_i^T x_i \vec{w}_j + \lambda \, sign(\vec{w}_j) \right) \right)$$

(c) Logistic Regression

$$L(\vec{w}) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \vec{w}^T \vec{x}_i))$$

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = \sum_{i=1}^{n} \frac{\exp(-y_i \vec{w}^T \vec{x}_i)'}{1 + \exp(-y_i \vec{w}^T \vec{x}_i)} = \sum_{i=1}^{n} \frac{-\vec{x}_i^{\mathrm{T}} \mathrm{y_i} \exp(-y_i \vec{w}^T \vec{x}_i)}{1 + \exp(-y_i \vec{w}^T \vec{x}_i)}$$

$$\vec{w}_{j+1} = \vec{w}_j - c \cdot \frac{\partial L(\vec{w}_j)}{\partial \vec{w}_j} = \vec{w}_j - c \cdot \left(\sum_{i=1}^{n} \frac{-\vec{x}_i^{\mathrm{T}} \mathrm{y_i} \exp(-y_i \vec{w}_j{}^T \vec{x}_i)}{1 + \exp(-y_i \vec{w}_j{}^T \vec{x}_i)} \right)$$

(d) Linear Support Vector Machine

$$L(\vec{w}) = C\sum_{i=1}^{n}\max\{1 - y_i\vec{w}^T\vec{x}_i, 0\} + \left|\left|\vec{w}\right|\right|_2^2$$

For each piece of the training data

$$\frac{\partial l_i(\vec{w})}{\partial \vec{w}} = \begin{cases} 0, & y_i\vec{w}^T x_i \geq 1 \\ -y_i x_i, & y_i\vec{w}^T x_i < 1 \end{cases}$$

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = \sum_{i=1}^{n}\frac{\partial l_i(\vec{w})}{\partial \vec{w}}$$

$$\vec{w}_{j+1} = \vec{w}_j - c\cdot\frac{\partial L(\vec{w}_j)}{\partial \vec{w}_j} = \vec{w}_j - c\cdot\left(\sum_{i=1}^{n}\frac{\partial l_i(\vec{w})}{\partial \vec{w}}\right)$$

Problem 2

(a) Objective function with new label

Because $y \in \{0,1\}$

$$P(y=1|\vec{w}^T\vec{x}) = \frac{1}{1+e^{-\vec{w}^T\vec{x}}}$$

$$P(y=0|\vec{w}^T\vec{x}) = 1 - P(y=1|\vec{w}^T\vec{x}) = \frac{1}{1+e^{\vec{w}^T\vec{x}}}$$

Then we have the loss function

$$L_0(\vec{w}) = -\log\left(\prod_{i=1}^{m} P(y_i|\vec{w}^T\vec{x}_i)\right) = -\sum_{i=1}^{n}\log\big(P(y_i|\vec{w}^T\vec{x}_i)\big)$$
$$= -\sum_{i=1}^{n}\big(y_i\log\big(sigm(\vec{w}^T\vec{x}_i)\big) + (1-y_i)\log\big(1-sigm(\vec{w}^T\vec{x}_i)\big)\big)$$

Because

$$\log\big(P(y_i|\vec{w}^T\vec{x}_i)\big) = \log\left(\left(\frac{1}{1+e^{-\vec{w}^T\vec{x_i}}}\right)^{y_i}\left(1-\frac{1}{1+e^{-\vec{w}^T\vec{x_i}}}\right)^{1-y_i}\right)$$
$$= y_i\log\left(\frac{1}{1+e^{-\vec{w}^T\vec{x_i}}}\right) + (1-y_i)\log\left(1-\frac{e}{1+e^{-\vec{w}^T\vec{x_i}}}\right)$$
$$= y_i\log\left(\frac{e^{\vec{w}^T\vec{x}_i}\left(1+e^{-\vec{w}^T\vec{x_i}}\right)}{1+e^{-\vec{w}^T\vec{x_i}}}\right) + \log\left(\frac{e}{1+e^{\vec{w}^T\vec{x}_i}}\right)$$
$$-\log\big(P(y_i|\vec{w}^T\vec{x}_i)\big) = -y_i\vec{w}^T\vec{x_i} + \log\left(\frac{1+e^{\vec{w}^T\vec{x}_i}}{e}\right)$$

We have

$$L_0(\vec{w}) = -\sum_{i=1}^{n}\log\big(P(y_i|\vec{w}^T\vec{x}_i)\big) = \sum_{i=1}^{n}\left(-y_i\vec{w}^T\vec{x_i} + \log\left(1+e^{\vec{w}^T\vec{x}_i}\right)\right)$$

For the logistic regression with $y \in \{-1,1\}$, its similar that

$$P(y_i|\vec{w}^T\vec{x}_i) = \left(\frac{1}{1+e^{-y_i\vec{w}^T\vec{x}_i}}\right)$$

With the same minimization process

$$L_{-1}(\vec{w}) = \log\left(\prod_{i=1}^{m} P(y_i|\vec{w}^T\vec{x}_i)\right) = -\sum_{i=1}^{n}\log\big(P(y_i|\vec{w}^T\vec{x}_i)\big) = \sum_{i=1}^{n}\log\left(1+e^{-y_i\vec{w}^T\vec{x}_i}\right)$$

When $y_i = 1$ it is clear that $L_0 = L_1$. When $y_i \neq 1$, for L_0 we have $y_i = 0$ and for L_{-1} we have $y_i = -1$

$$0 + \log\left(1+e^{\vec{w}^T\vec{x}_i}\right) = \log(1+e^{-(-1)\vec{w}^T\vec{x}_i}) \rightarrow L_0 = L_{-1}$$

Therefore, the logistic regression loss function L_0 for $y_i \in \{0,1\}$ is equal to the logistic regression loss function L_{-1} for $y \in \{-1,1\}$.

(b) Gradient

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = \left(\sum_{i=1}^{n}\left(-y_i \vec{w}^T \vec{x} + \log\left(1 + e^{\vec{w}^T \vec{x}}\right)\right)\right)' = \sum_{i=1}^{n}\left(-y_i \vec{x_i} + \frac{e^{\vec{w}^T \vec{x}_i}}{1 + e^{\vec{w}^T \vec{x}_i}} \cdot \vec{x}_i\right)$$

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = \sum_{i=1}^{n} -\left(y_i - \frac{e^{\vec{w}^T \vec{x}_i}}{1 + e^{\vec{w}^T \vec{x}_i}}\right)\vec{x}_i = -\sum_{i=1}^{n}\left(y_i - sigm(\vec{w}^T \vec{x}_i)\right)\vec{x}_i$$

(c) Hessian Matrix

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = -\sum_{i=1}^{n}\left(y_i - sigm(\vec{w}^T \vec{x}_i)\right)\vec{x}_i$$

$$\frac{\partial L(\vec{w})}{\partial \vec{w}\partial \vec{w}} = -\sum_{i=1}^{n}\left(y_i - sigm(\vec{w}^T \vec{x}_i)\right)\vec{x}_i = \sum_{i=1}^{n} \vec{x}_i^T \left(sigm(\vec{w}^T \vec{x}_i)\left(1 - sigm(\vec{w}^T \vec{x}_i)\right)\right)\vec{x}_i$$

Because we have the diagonal matrix W such that

$$W_{ii} = sigm(\vec{w}^T \vec{x}_i)\left(1 - sigm(\vec{w}^T \vec{x}_i)\right)$$

$$H = \frac{\partial L(\vec{w})}{\partial \vec{w}\partial \vec{w}} = \sum_{i=1}^{n} \vec{x}_i \left(sigm(\vec{w}^T \vec{x}_i)\left(1 - sigm(\vec{w}^T \vec{x}_i)\right)\right)\vec{x}_i = X^T W X$$

Because sigmoid function ranges from 0 to 1

$$W_{iimax} = 0.25 \leftrightarrow sigm(\vec{w}^T \vec{x}_i) = 1 - sigm(\vec{w}^T \vec{x}_i)$$

$$W_{iimin} = 0 \leftrightarrow sigm(\vec{w}^T \vec{x}_i) = 0 \; or \; 1$$

(d) Convex

$$H = \sum_{i=1}^{n} \vec{x}_i^T \left(sigm(\vec{w}^T \vec{x}_i)\left(1 - sigm(\vec{w}^T \vec{x}_i)\right)\right)\vec{x}_i$$

Because $sigm(\vec{w}^T \vec{x}_i) \in [0,1]$ and $1 - sigm(\vec{w}^T \vec{x}_i) \in [0,1]$

$$sigm(\vec{w}^T \vec{x}_i)\left(1 - sigm(\vec{w}^T \vec{x}_i)\right) \geq 0$$

That is, the Hessian is the linear combination of the product of a squared term

$$H = \sum_{i=1}^{n} \vec{x}_i^T \left(sigm(\vec{w}^T \vec{x}_i)\left(1 - sigm(\vec{w}^T \vec{x}_i)\right)\right)\vec{x}_i \geq 0$$

Therefore, the negative log-likelihood is convex because the Hessian is positive semi definite

(e) Newton update

$$\vec{w}_{new} = \vec{w}_{old} + (X^T W X)^{-1} X^T \left(y - sigm\left(\vec{w}_{old}^T X\right)\right)$$
$$= (X^T W X)^{-1} X^T W \left(X^T \vec{w}_{old} + W^{-1}\left(y - sigm\left(\vec{w}_{old}^T X\right)\right)\right)$$

Because we have a substitution

$$\vec{z}_i = x_i{}^T \vec{w} + \frac{1}{W_{ii}}\left(y_i - sigm(\vec{w}^T x_i)\right)$$

$$\vec{w}_{new} = (X^T W X)^{-1} X^T W \left(X^T \vec{w}_{old} + W^{-1}\left(y - sigm\left(\vec{w}_{old}^T X\right)\right)\right) = (X^T W X)^{-1} X^T W \vec{z}$$

Problem 3

(a) Weighted ridge regression

$$L(\vec{w}) = \sum_{i=1}^{n} p_i \, (\vec{w}^T \vec{x}_i - y_i)^2 + \lambda \vec{w}^T \vec{w} = (\vec{w}^T X - Y)^T P (\vec{w}^T X - y) + \lambda \vec{w}^T \vec{w}$$

(b) $\vec{w}$ solution

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = \sum_{i=1}^{n} P\left(y_i^T y_i - 2\vec{w}^T \vec{x}_i{}^T y_i + \vec{w}^T \vec{x}_i^T \vec{x}_i \vec{w} \right)' + (\vec{w}^T \lambda I \vec{w})' = P \sum_{i=1}^{n} (-2\vec{x}_i{}^T y_i + 2\vec{x}_i^T \vec{x}_i \vec{w}) + 2\lambda \vec{w}$$

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = -X^T P Y + (X^T P X + \lambda)\vec{w}$$

Set the gradient to zero

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = -X^T P Y + (X^T P X + \lambda)\vec{w} \rightarrow X^T P Y = (X^T P X + \lambda)\vec{w}$$

$$\vec{w} = (X^T P X + \lambda)^{-1} X^T P Y$$

(c) Iterative reweighted least square

With λ set to 0

$$\vec{w} = (X^T P X + 0)^{-1} X^T P Y = (X^T P X)^{-1} X^T P Y$$

It is called iteratively reweighted least square because the weight matrix is being updated iteratively so that the instance on which the learner makes mistakes gets higher weight. Therefore the algorithm can find the most likely estimation of the regression.