CS 189 Introduction to Machine Learning Spring2019 Homework 4

Q1 Logistic Regression with Newton's Method

part 1

$$\Delta_w J = 2\lambda w + X^\intercal (s - y)$$

part 2

$$\Delta_w^2 J = 2\lambda I + X^\intercal \Omega X$$

where $\Omega = \text{diag}(z)\text{diag}(1 - z)$

part 3

$$w_{t+1} \leftarrow w_t - (2\lambda I + X^\intercal \Omega X)^{-1}(2\lambda w_t + X^\intercal (s - y))$$

part 4

```python
def sigmoid(z):
    return 1 / (1 + np.exp(-z))

def fit_once(X, y, w, l):
    z = sigmoid(X @ w)
    Omega = np.diagflat(z * (1 - z))
    DF = X.T @ (z - y) + 2 * l * w
    D2F = (X.T @ Omega @ X) + 2 * l * np.identity(w.shape[0])
    e = - np.linalg.inv(D2F) @ DF
    return z, w + e

if __name__ == "__main__":
    X = np.asarray([0, 3, 1, 3, 0, 1, 1, 1]).reshape([4, 2]).astype('float64')
    y = np.asarray([1, 1, 0, 0]).reshape([4, 1]).astype('int32')

    n = X.shape[0]  # n = 4
    d = X.shape[1]  # d = 2

    X = np.concatenate([X, np.ones([n, 1])], axis=1)
    w0 = np.asarray([-2, 1, 0]).reshape([d + 1, 1]).astype('float64')
    l = 0.07

    s0, w1 = fit_once(X, y, w0, l)
    s1, w2 = fit_once(X, y, w1, l)

    np.set_printoptions(precision=4)
    np.set_printoptions(suppress=True)
    print('s0:', s0)
    print('w1:', w1)
    print('s1:', s1)
    print('w2:', w2)

    # Results:
    # s0: [[0.9526]
    #     [0.7311]
    #     [0.7311]
    #     [0.2689]]
    # w1: [[-0.3868]
    #     [ 1.4043]
    #     [-2.2842]]
    # s1: [[0.8731]
    #     [0.8238]
    #     [0.2932]
    #     [0.2198]]
    # w2: [[-0.5122]
    #     [ 1.4527]
    #     [-2.1627]]
```

Q2 l_1- and l_2-Regularization

part 1

$$J(w) = y^\intercal y + \sum_{i=1}^{d}(\lambda|w_i| + nw_i^2 - 2y^\intercal X_{*i}w_i)$$

so, $g(y) = y^\intercal y$, and $f(X_{*i}, w_i, y, \lambda) = \lambda|w_i| + nw_i^2 - 2y^\intercal X_{*i}w_i$

part 2

If $w_i^* > 0$, then $f(X_{*i}, w_i, y, \lambda) = \lambda w_i + nw_i^2 - 2y^\intercal X_{*i}w_i$, and $\Delta_{w_i} f(X_{*i}, w_i, y, \lambda) = \lambda - 2y^\intercal X_{*i} + 2nw_i$.

$$\Delta_{w_i} f(X_{*i}, w_i, y, \lambda) = 0 \Rightarrow w_i^* = \frac{1}{n}(y^\intercal X_{*i} - \lambda/2)$$

part 3

Similar to part 2,

$$\Delta_{w_i} f(X_{*i}, w_i, y, \lambda) = 0 \Rightarrow w_i^* = \frac{1}{n}(y^\intercal X_{*i} + \lambda/2)$$

part 4

w_i^* can not be greater than 0 if $\frac{1}{n}(y^\intercal X_{*i} - \lambda/2) \le 0$, i.e. $2y^\intercal X_{*i} \le \lambda$; w_i^* can not be less than 0 if $\frac{1}{n}(y^\intercal X_{*i} + \lambda/2) \ge 0$, i.e. $2y^\intercal X_{*i} \ge -\lambda$. w_i^* is zero if both are true, i.e., $-\lambda \le 2y^\intercal X_{*i} \le \lambda$.

part 5

$f'(X_{*i}, w_i, y, \lambda) = \lambda w_i^2 + nw_i^2 - 2y^\intercal X_{*i}w_i$.
Setting the derivative to 0 yields

$$w_i^* = \frac{y^\intercal X_{*i}}{n + \lambda}$$

Therefore, w_i^* is 0 if $y^\intercal X_{*i} = 0$. It is a much stronger condition than $|y^\intercal X_{*i}| < \lambda/2$ in Lasso. This shows why l_1-regularization encourages sparsity.

Q3 Regression and Dual Solutions

part 1

$$\Delta|w|^4 = (w^\intercal w)^4 = 4(w^\intercal w)w$$

Let $l(w) = |Xw - y|^2$, then $\Delta_w l = 2X^\intercal Xw - 2X^\intercal y$

$$\Delta_w|Xw - y|^4 = \Delta_w l^2 = 2l(\Delta_w l) = 4|Xw - y|^2(X^\intercal Xw - X^\intercal y)$$

part2

Let $J(w) = |Xw - y|^4 + \lambda|w|^2$. Then we have

$$\Delta_w J = 4|Xw - y|^2(X^\intercal Xw - X^\intercal y) + 2\lambda w$$

Setting $\Delta_w J = 0$ gives

$$w^* = \frac{2|Xw^* - y|^2}{\lambda}X^\intercal(y - Xw^*) = X^\intercal a$$

where $a = \frac{2|Xw^* - y|^2}{\lambda}(y - Xw^*)$

To show that the optimum w^* is unique, we compute the Hessian of the objective, note that $\Delta_w^2 l = 2X^\intercal X$

$$\begin{aligned}\Delta_w^2 J &= 4\Delta_w l(\frac{1}{2}\Delta_w l^\intercal) + 4l(\frac{1}{2}\Delta_w^2 l) + 2\lambda I_d \\ &= 2(\Delta_w l)(\Delta_w l)^\intercal + 2l(\Delta_w^2 l) + 2\lambda I_d\end{aligned}$$

We claim that $\Delta_w^2 J$ is positive definite.

Proof: For any $z \in \mathbb{R}^d$ and $z \neq \mathbf{0}$, $z^\intercal(\Delta_w^2 J)z = 2|(\Delta_w l)^\intercal z|^2 + 4l|Xz|^2 + 2\lambda|z|^2 > 0$, since $|(\Delta_w l)^\intercal z|^2 \geq 0, l \geq 0, |Xz|^2 \geq 0, \lambda > 0$, and $|z|^2 > 0$. **QED.**

Therefore, $J(w)$ is strict convex, so the optimum w^* is unique.

part3

Suppose $w^{*\prime} = w^* + v$, where $w^* = X^\intercal a$ and v is in the null space of X, i.e. $Xv = 0$. Note that $w^{*\intercal}v = 0$.

Let $J(w) = \frac{1}{n}\sum_{i=1}^{n} L(w^\intercal X_i, y_i) + \lambda |w|^2$, then

$$\begin{aligned} J(w^{*\prime}) &= \frac{1}{n}\sum_{i=1}^{n} L((w^* + v)^\intercal X_i, y_i) + \lambda |w^* + v|^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} L(w^{*\intercal} X_i, y_i) + \lambda |w^*|^2 + \lambda |v|^2 \end{aligned}$$

In order to make J minimized, v has to be $\mathbf{0}$, i.e. w^* has the form $w^* = \sum_{i=1}^{n} a_i X_i$. This holds for L being both convex and non-convex.

Q5 Real World Spam Classification

Use a binary feature indicating whether a time stamp is "close to midnight" or not instead of the number of milliseconds since previous midnight. Then use the linear SVM