1 Bayes Optimal Classification (20 Points) (Yan)

1. We can write

$$
\begin{aligned}
\underset{f}{\operatorname{argmin}}\, \mathbb{E}\ell_{\alpha,\beta}(f(x),y) &= \underset{f}{\operatorname{argmin}}\, \mathbb{E}_{X,Y}\left[\alpha\mathbf{1}\left\{f(X)=1,Y=0\right\}+\beta\mathbf{1}\left\{f(X)=0,Y=1\right\}\right] \\
&= \underset{f}{\operatorname{argmin}}\, \mathbb{E}_{X}\left[\mathbb{E}_{Y\mid X}\left[\alpha\mathbf{1}\left\{f(X)=1,Y=0\right\}+\beta\mathbf{1}\left\{f(X)=0,Y=1\right\}\right]\right] \\
&= \underset{f}{\operatorname{argmin}}\, \mathbb{E}_{X}\left[\int_y \alpha\mathbf{1}\left\{f(X)=1,Y=0\right\}+\beta\mathbf{1}\left\{f(X)=0,Y=1\right\}dP(y\mid x)\right] \\
&= \underset{f}{\operatorname{argmin}} \int_x \left[\alpha\mathbf{1}\left\{f(x)=1\right\}P(y=0\mid x)+\beta\mathbf{1}\left\{f(x)=0\right\}P(y=1\mid x)\right]dP(x)
\end{aligned}
$$

We may minimize he integrand at each x by taking

$$
f(x)=\begin{cases}1 & \beta P(y=1\mid x)\geq \alpha P(y=0\mid x)\\ 0 & \alpha P(y=0\mid x)>\beta P(y=1\mid x)\end{cases}
$$

2. Notice that

$$
\begin{aligned}
\mathbb{E}\ell_{\alpha,\beta}(f(x),y) &= \alpha P(f(x)=1,y=0)+\beta P(f(x)=0,f=1) \\
&= \alpha P(f(x)=1\mid y=0)P(y=0)+\beta P(f(x)=0\mid f=1)P(y=1)
\end{aligned}
$$

which is same as the minimizer of the given risk R if $\alpha=\frac{1}{P(y=0)}$ and $\beta=\frac{1}{P(y=1)}$.

3. Notice that since $Y\sim \operatorname{Ber}\left(\frac{1}{2}\right)$, we have $P(Y=1)=P(Y=0)=\frac{1}{2}$.

$$
\begin{aligned}
f^\star(x) &= \underset{y}{\operatorname{argmax}}\, P(Y=y\mid X=x)=\underset{y}{\operatorname{argmax}}\, P(X=x\mid Y=y)P(Y=y) \\
&= \underset{y}{\operatorname{argmax}}\, P(X=x\mid Y=y)
\end{aligned}
$$

Therefore, $f^\star(1)=1$ since $p=P(X=1\mid Y=1)>P(X=1\mid Y=0)=q$, and $f^\star(0)=0$ since $1-p=P(X=0\mid Y=1)<P(X=0\mid Y=0)=1-q$. Hence, $f^\star(X)=X$. The risk is $R^\star=P(f\star(X)\neq Y)=P(X\neq Y)$.

$$
R^\star=P(Y=1)P(X=0\mid Y=1)+P(Y=0)P(X=1\mid Y=0)=\frac{1}{2}\cdot(1-p)+\frac{1}{2}\cdot q
$$

4. Figure 1 is a sample plot for this problem.

2 Regularized Linear Regression Using Lasso (20 Points) (Yan)

1. We can expand the equation as

$$
\begin{aligned}
&\frac{1}{2}(y^\top y-2y^\top X\mathbf{w}+\mathbf{w}^\top X^\top X\mathbf{w})+\lambda\left\|\mathbf{w}\right\|_1 \\
=&\frac{1}{2}y^\top y+\sum_{i=1}^{d}-y^\top X_{.i}\mathbf{w}_i+\frac{1}{2}\mathbf{w}_i^2+\lambda|\mathbf{w}_i|
\end{aligned}
$$

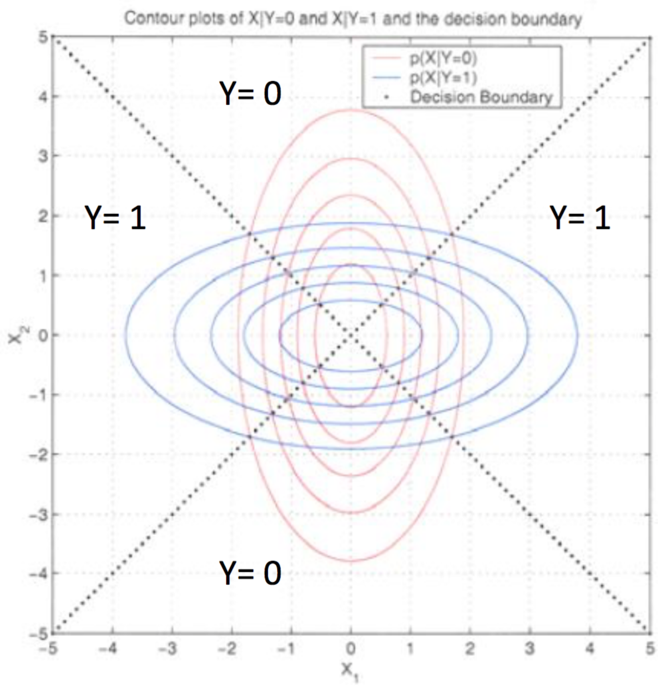


Figure 1: Sample plot for problem 1-4

2. If $\mathbf{w}_i^\star > 0$, then we want to maximize

$$-y^\top X_{.i}\mathbf{w}_i + \frac{1}{2}\mathbf{w}_i^2 + \lambda\mathbf{w}_i$$

   Take the derivative and equate to 0, we have:

$$\mathbf{w}_i^\star = y^\top X_{.i} - \lambda$$

3. If $\mathbf{w}_i^\star < 0$, then we want to maximize

$$-y^\top X_{.i}\mathbf{w}_i + \frac{1}{2}\mathbf{w}_i^2 - \lambda\mathbf{w}_i$$

   Take the derivative and equate to 0, we have:

$$\mathbf{w}_i^\star = y^\top X_{.i} + \lambda$$

4. From the previous questions, we knoww $w_i^\star = 0$ if none of the above conditions hold, that is

$$y^\top X_{.i} - \lambda \leq 0 \quad y^\top X_{.i} + \lambda \geq 0$$

   Combining them, we get

$$-\lambda \leq y^\top X_{.i} \leq \lambda$$

5. If the lasso is replaced by $\frac{1}{2}\lambda\left\|\mathbf{w}\right\|_2^2$, the optimization problem regarding $\mathbf{w}_i$ is given by

$$-y^\top X_{.i}\mathbf{w}_i + \frac{1}{2}\mathbf{w}_i^2 + \frac{1}{2}\lambda\mathbf{w}_i^2$$

Take the derivative and equate to 0, we have:

$$\mathbf{w}_i^\star = \frac{y^\top X_{.i}}{1+\lambda}$$

It is equal to 0 if $y^\top X_{.i} = 0$ or λ goes to infinity. In contrast, $\mathbf{w}_i^\star = 0$ when $\left|y^\top X_{.i}\right| < \lambda$ in Lasso regression. This is why the L1 norm regularization encourages sparsity.

3 Multinomial Logistic Regression (20 Points) (Yan)

The solution for this question absorbs the intercept term w_{c0} into the vector of $\mathbf{w}_c$.

1. When $C = 2$,

$$\begin{aligned}
p(y=1 \mid \mathbf{x}, W) &= \frac{\exp\left(w_{10} + \mathbf{w}_1^\top \mathbf{x}\right)}{\exp\left(w_{10} + \mathbf{w}_1^\top \mathbf{x}\right) + \exp\left(w_{20} + \mathbf{w}_2^\top \mathbf{x}\right)} = \frac{1}{1 + \exp\left(w_{20} - w_{10} + (\mathbf{w}_2 - \mathbf{w}_1)^\top \mathbf{x}\right)} \\
p(y=2 \mid \mathbf{x}, W) &= \frac{\exp\left(w_{20} + \mathbf{w}_2^\top \mathbf{x}\right)}{\exp\left(w_{10} + \mathbf{w}_1^\top \mathbf{x}\right) + \exp\left(w_{20} + \mathbf{w}_2^\top \mathbf{x}\right)} = \frac{\exp\left(w_{20} - w_{10} + (\mathbf{w}_2 - \mathbf{w}_1)^\top \mathbf{x}\right)}{1 + \exp\left(w_{20} - w_{10} + (\mathbf{w}_2 - \mathbf{w}_1)^\top \mathbf{x}\right)}
\end{aligned}$$

   This is equivalent with logistic regression that has weights $(w_{20} - w_{10}, \mathbf{w}_2 - \mathbf{w}_1)$.

2. Let $\mu_{ic} = P(y_i = c \mid \mathbf{x}_i, W)$, $y_{ic} = \mathbf{1}\{y_i = c\}$.

   (a)

$$\ell(W) = \log \prod_{i=1}^n \prod_{c=1}^C \mu_{ic}^{y_{ic}} = \sum_{i=1}^n \sum_{c=1}^C y_{ic} \log \mu_{ic} = \sum_{i=1}^n \left(\sum_{c=1}^C y_{ic} \mathbf{w}_c^\top \mathbf{x}_i - \log \sum_{c'=1}^C \exp\left(\mathbf{w}_{c'}^\top \mathbf{x}_i\right) \right)$$

   (b)

$$\begin{aligned}
g_c(W) &= \frac{\partial}{\partial \mathbf{w}_c} \sum_{i=1}^n \left(\sum_{c=1}^C y_{ic} \mathbf{w}_c^\top \mathbf{x}_i - \log \sum_{c'=1}^C \exp\left(\mathbf{w}_{c'}^\top \mathbf{x}_i\right) \right) \\
&= \sum_{i=1}^n \left(\frac{\partial}{\partial \mathbf{w}_c} \sum_{c=1}^C y_{ic} \mathbf{w}_c^\top \mathbf{x}_i - \frac{\partial}{\partial \mathbf{w}_c} \log \sum_{c'=1}^C \exp\left(\mathbf{w}_{c'}^\top \mathbf{x}_i\right) \right) \\
&= \sum_{i=1}^n \left(y_{ic} \mathbf{x}_i - \frac{\frac{\partial}{\partial \mathbf{w}_c} \sum_{c'=1}^C \exp\left(\mathbf{w}_{c'}^\top \mathbf{x}_i\right)}{\sum_{c'=1}^C \exp\left(\mathbf{w}_{c'}^\top \mathbf{x}_i\right)} \right) \\
&= \sum_{i=1}^n \left(y_{ic} \mathbf{x}_i - \frac{\exp\left(\mathbf{w}_c^\top \mathbf{x}_i\right) \mathbf{x}_i}{\sum_{c'=1}^C \exp\left(\mathbf{w}_{c'}^\top \mathbf{x}_i\right)} \right) \\
&= \sum_{i=1}^n (y_{ic} - \mu_{ic}) \mathbf{x}_i
\end{aligned}$$

(c) $\delta_{cc'}$ denotes the Dirac delta function and is equal to one if $c = c'$ and zero otherwise.

$$
\begin{aligned}
H_{c,c'}(W) &= \frac{\partial}{\partial \mathbf{w}_c} g_{c'}(W) \\
&= \frac{\partial}{\partial \mathbf{w}_c} \sum_{i=1}^{n} \left(y_{ic'} \mathbf{x}_i - \frac{\exp\left(\mathbf{w}_{c'}^\top \mathbf{x}_i\right) \mathbf{x}_i}{\sum_{c''=1}^{C} \exp\left(\mathbf{w}_{c''}^\top \mathbf{x}_i\right)} \right) \\
&= -\sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{w}_c} \frac{\exp\left(\mathbf{w}_{c'}^\top \mathbf{x}_i\right) \mathbf{x}_i}{\sum_{c''=1}^{C} \exp\left(\mathbf{w}_{c''}^\top \mathbf{x}_i\right)} \\
&= -\sum_{i=1}^{n} \frac{\delta_{cc'} \exp\left(\mathbf{w}_{c'}^\top \mathbf{x}_i\right) \mathbf{x}_i \mathbf{x}_i^\top \sum_{c''=1}^{C} \exp\left(\mathbf{w}_{c''}^\top \mathbf{x}_i\right) - \exp\left(\mathbf{w}_{c'}^\top \mathbf{x}_i\right) \mathbf{x}_i \mathbf{x}_i^\top \exp\left(\mathbf{w}_{c}^\top \mathbf{x}_i\right)}{\left(\sum_{c''=1}^{C} \exp\left(\mathbf{w}_{c''}^\top \mathbf{x}_i\right)\right)^2} \\
&= -\sum_{i=1}^{n} \left(\delta_{cc'} \mu_{ic} - \mu_{ic'} \mu_{ic}\right) \mathbf{x}_i \mathbf{x}_i^\top \\
&= \sum_{i=1}^{n} \mu_{ic} \left(\mu_{ic'} - \delta_{cc'}\right) \mathbf{x}_i \mathbf{x}_i^\top
\end{aligned}
$$

4 Perceptron Mistake Bounds (20 Points) (Xun)

1. Expand $\mathbf{w}^{(t)}$:

$$
\begin{aligned}
\langle \mathbf{w}^{(t)}, \mathbf{w} \rangle &= \langle \mathbf{w}^{(t-1)} + y^{(t)} \mathbf{x}^{(t)}, \mathbf{w} \rangle && (1) \\
&= \langle \mathbf{w}^{(t-1)}, \mathbf{w} \rangle + \langle y^{(t)} \mathbf{x}^{(t)}, \mathbf{w} \rangle && (2) \\
&\geq \langle \mathbf{w}^{(t-1)}, \mathbf{w} \rangle + \gamma, && (3)
\end{aligned}
$$

where the inequality holds due to the separability assumption. Expand recursively until $\mathbf{w}^{(0)}$, we get the desired inequality.

2. Similarly,

$$
\begin{aligned}
\|\mathbf{w}^{(t)}\|_2^2 &= \|\mathbf{w}^{(t-1)}\|_2^2 + 2 \cdot \langle y^{(t)} \mathbf{x}^{(t)}, \mathbf{w}^{(t-1)} \rangle + \|\mathbf{x}^{(t)}\|_2^2 && (4) \\
&\leq \|\mathbf{w}^{(t-1)}\|_2^2 + M^2, && (5)
\end{aligned}
$$

where we use the fact that $(\mathbf{x}^{(t)}, y^{(t)})$ is a misclassified example, therefore $\left\langle y^{(t)} \mathbf{x}^{(t)}, \mathbf{w}^{(t-1)} \right\rangle \leq 0$. Result follows by recursively expand the inequality until $\mathbf{w}^{(0)}$.

3. Square both sides of the first inequality and apply Cauchy-Schwartz:

$$
t^2 \gamma^2 \leq \langle \mathbf{w}^{(t)}, \mathbf{w} \rangle^2 \leq \|\mathbf{w}^{(t)}\|_2^2 \cdot \|\mathbf{w}\|_2^2 \leq t M^2. \qquad (6)
$$

4. False. If data is linearly separable by a margin, then there are infinitely many classifiers that achieve zero error. Depending on the ordering of the data, Perceptron may stop at any of the hyperplanes.

   If a stronger argument is favored, we can easily construct a 2d example that achieves zero error even at $\mathbf{w}^{(0)}$, which may not have margin γ.

5 Logistic Regression for Image Classification (20 Points) (Xun)

5.1 Exploring the data

1. Each image is sized 28×28, so expressed as a 784×1 vector.

```
        660        898.543      0.000102349              18.8
        661        898.451       0.00010217             18.82
        662         898.36      0.000101992           18.8401
        663        898.268      0.000101814           18.8601
        664        898.177      0.000101636           18.8801
        665        898.086       0.00010146           18.9001
        666        897.995      0.000101283           18.9201
        667        897.904      0.000101107           18.9401
        668        897.814      0.000100932           18.9601
        669        897.723      0.000100757           18.9801
        670        897.633      0.000100583           19.0001
        671        897.543      0.000100409           19.0201
        672        897.453      0.000100236           19.0401
        673        897.363      0.000100063             19.06
        674        897.273      9.98907e-05             19.08
Training accuracy: 97.8551%
Test accuracy: 96.8246%
```

(a)

```
>> oracle_lr_test

err =

   1.0e-06 *

    0.2910
    0.2676
    0.2817
    0.2862
    0.2910
    0.2724
    0.2893
    0.2708
    0.2828
    0.2760

Pass!
```

(b)

Figure 2: Results of binary logistic regression without regularization.

```
        310        1009.74      0.000108976           9.54126
        311        1009.63      0.000108376            9.5546
        312        1009.53       0.00010778            9.5679
        313        1009.42      0.000107189           9.58119
        314        1009.31      0.000106602           9.59444
        315         1009.2      0.000106019           9.60767
        316         1009.1       0.00010544           9.62087
        317        1008.99      0.000104865           9.63405
        318        1008.89      0.000104294            9.6472
        319        1008.78      0.000103727           9.66033
        320        1008.68      0.000103164           9.67343
        321        1008.57      0.000102605            9.6865
        322        1008.47       0.00010205           9.69955
        323        1008.37      0.000101499           9.71257
        324        1008.27      0.000100951           9.72557
        325        1008.16      0.000100408           9.73854
        326        1008.06       9.9868e-05           9.75149
Training accuracy: 97.7466%
Test accuracy: 96.9254%
```

(a)

```
>> oracle_lr_test

err =

   1.0e-06 *

    0.2776
    0.3032
    0.2956
    0.2936
    0.2838
    0.2869
    0.2837
    0.2952
    0.2922
    0.2997

Pass!
```

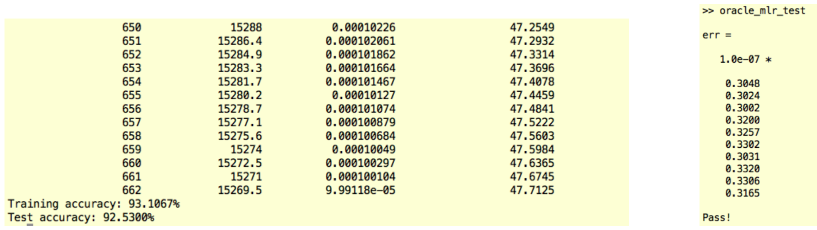
(b)

Figure 3: Results of binary logistic regression with regularization.

2. 0 to 9.
3. 0 to 1.
4. 3.5698 and 17.1790.
5. Nonzero fraction is 0.1912. Depending on the sparsity level, you could either argue it is sparse or dense.
6. Close to uniform.

5.2 Binary logistic regression

1. See the reference code.
2. See the reference code.
3. See the reference code. Figure 2(b) shows the result.
4. See the reference code.
5. See the reference code.
6. Figure 2(a) and (b) show the result produced by the reference code.
7. Figure 3(a) and (b) show the result produced by the reference code.



(a) (b)

Figure 4: Results of multiclass logistic regression.

5.3 Multiclass logistic regression

1. See the reference code.
2. See the reference code.
3. See the reference code.
4. Figure 4(a) and (b) show the result produced by the reference code.