TSINGHUA UNIVERSITY, DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

Machine Learning HW1

计算机系　刘泓尊　2022210866

2022 年 10 月 16 日

1 Collaborators and Sources

I finished this assignment independently but refered to some slides from CMU indeed.

References

1 Support Vector Machines, Kernel SVM - Carnegie Mellon University

2 Soft margin SVM - Carnegie Mellon University

2 Maximum Likelihood Estimators

Problem 1

1.

$$
\begin{aligned}
l &= logL \\
&= log\prod_{i=1}^{N} p(x_i|\mu, \Sigma) \\
&= log\prod_{i=1}^{N} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \\
&= \sum_{i=1}^{N}\left(-\frac{d}{2}log(2\pi) - \frac{1}{2}log|\Sigma| - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right) \\
&= -\frac{Nd}{2}log(2\pi) - \frac{N}{2}log|\Sigma| + \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)
\end{aligned}
$$

Deriving μ If matrix A is symmetric, then

$$\frac{\partial w^T A w}{\partial w} = 2Aw$$

Thus, we let

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{N} \Sigma^{-1}(\mu - x_i) = 0$$

Since Σ is positive definite, we have

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Deriving Σ We know that

$$tr(ACB) = tr(CAB) = tr(BCA)$$

and

$$\frac{\partial tr(AB)}{A} = B^T$$

thus we have

$$x^T A x = tr(x^T A x) = tr(x^T x A)$$

so

$$\frac{\partial x^T A x}{\partial A} = \frac{\partial x^T x A}{\partial A} = x x^T$$

$$\frac{\partial log|A|}{\partial A} = A^{-T}$$

Thus, we let $\frac{\partial l}{\partial \Sigma} = 0$, since A is symmetric, we have

$$\begin{aligned}
\frac{\partial l}{\partial \Sigma} &= \frac{\partial}{\partial \Sigma}\left(C + \frac{N}{2} log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right) \\
&= \frac{N}{2}\Sigma^T - \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T \\
&= \frac{N}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T \\
&= 0
\end{aligned}$$

so

$$\hat{\Sigma}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

In conclusion

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

$$\hat{\Sigma}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{ML})(x_i - \hat{\mu}_{ML})^T \tag{2}$$

2.

$$E\left[\hat{\mu}_{ML}\right] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \mu$$

$$
\begin{aligned}
E\left[\hat{\Sigma}_{ML}\right] &= \frac{1}{N}E\left[\sum_{i=1}^{N}(x_i-\hat{\mu}_{ML})(x_i-\hat{\mu}_{ML})^T\right] \\
&= \frac{1}{N}E\left[\sum_{i=1}^{N}(x_i x_i^T - x_i\hat{\mu}_{ML}^T - \hat{\mu}_{ML}x_i^T + \hat{\mu}_{ML}\hat{\mu}_{ML}^T)\right] \\
&= \frac{1}{N}E\left[\sum_{i=1}^{N}x_i x_i^T - N\hat{\mu}_{ML}\hat{\mu}_{ML}^T\right] \\
&= \frac{1}{N}E[\sum_{i=1}^{N}x_i x_i^T] - E[\hat{\mu}_{ML}\hat{\mu}_{ML}^T] \\
&= \frac{1}{N}\left(\sum_{i=1}^{N}\left[\Sigma(x_i)+E(x_i)E(x_i^T)\right]\right) - \left(\Sigma(\hat{\mu}_{ML})+E(\hat{\mu}_{ML})E(\hat{\mu}_{ML}^T)\right) \\
&= \frac{1}{N}\left(\sum_{i=1}^{N}(\Sigma+\mu\mu^T)\right) - \left(\frac{1}{N}\Sigma+\mu\mu^T\right) \\
&= \frac{N-1}{N}\Sigma
\end{aligned}
$$

In conclusion

$$E\left[\hat{\mu}_{ML}\right] = \mu \tag{3}$$

$$E\left[\hat{\Sigma}_{ML}\right] = \frac{N-1}{N}\Sigma \tag{4}$$

So $E\left[\hat{\mu}_{ML}\right]$ is an unbiased estimate, but $E\left[\hat{\Sigma}_{ML}\right]$ is a biased estimate.

3.

$$
\begin{aligned}
E[\|\hat{\mu}_{ML}-\mu\|^2] &= E[(\hat{\mu}_{ML}-\mu)^T(\hat{\mu}_{ML}-\mu)] \\
&= E[\hat{\mu}_{ML}^T\hat{\mu}_{ML}] - \mu^T\mu
\end{aligned}
$$

$$
\begin{aligned}
E[\hat{\mu}_{ML}^T\hat{\mu}_{ML}] &= E[Tr(\hat{\mu}_{ML}\hat{\mu}_{ML}^T)] \\
&= Tr\left(E[\hat{\mu}_{ML}\hat{\mu}_{ML}^T)]\right) \\
&= Tr\left(\frac{1}{N}\Sigma+\mu\mu^T\right) \\
&= \frac{1}{N}Tr\left(\Sigma\right)+\mu^T\mu
\end{aligned}
$$

So

$$E[\|\hat{\mu}_{ML}-\mu\|^2] = \frac{Tr\Sigma}{N} \tag{5}$$

3 Kernel SVM

Problem 2

1.

$$
\begin{aligned}
k(x,y) &= (1+xy)^n \\
&= \binom{n}{0} + \binom{n}{1}xy + \binom{n}{2}x^2y^2 + \cdots + \binom{n}{n}x^ny^n
\end{aligned}
$$

let

$$\phi(x) = \left(\sqrt{\binom{n}{0}}, \sqrt{\binom{n}{1}}x, \cdots, \sqrt{\binom{n}{n}}x^n \right) \in R^{n+1} \tag{6}$$

we have

$$k(x,y) = (1+xy)^n = \phi(x)^T\phi(y)$$

so $k(x,y) = (1+xy)^n$ is a kernel on $X = R$.

2.

Assume $k(x,y) = xy - 1$ is kernel function on $X = R$, then there exsits a function $\phi(\dot{)}$ such that $k(x,y) = \phi(x)^T\phi(y)$.

So for any $x \in R$, we have $k(x,x) = \phi(x)^T\phi(x) \geq 0$

Let $x = y = 0 \in R$, we have $k(0,0) = -1 < 0$.

We got a contradiction, so $k(x,y) = xy - 1$ is not a kernel function on $X = R$.

Problem 3

1.

$$\begin{aligned} L &= \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i(-\xi_i) + \sum_{i=1}^{N}\beta_i\left(1 - \xi_i - y_i w^T\phi(x_i)\right) \\ &= \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{N}(1-\alpha_i)\xi_i + \sum_{i=1}^{N}\beta_i\left(1 - \xi_i - y_i w^T\phi(x_i)\right) \\ s.t. \quad & \alpha_i \geq 0 \\ & \beta_i \geq 0 \end{aligned} \tag{7}$$

2.

$$\min_{w,\xi_i}\max_{\alpha_i,\beta_i} L$$

we let

$$\begin{aligned} \frac{\partial L}{\partial w} &= \lambda w - \sum_{i=1}^{N}\beta_i y_i\phi(x_i) = 0 \\ \frac{\partial L}{\partial \xi_i} &= (1-\alpha_i) - \beta_i = 0 \end{aligned}$$

we have

$$w = \frac{1}{\lambda}\sum_{i=1}^{N}\beta_i y_i\phi(x_i) \tag{8}$$

$$1 - \alpha_i = \beta_i \tag{9}$$

so $0 \leq \beta_i \leq 1$, then we have

$$\begin{aligned}
\min_{w,\xi_i} \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{N} \xi_i &= \min_{w,\xi_i} \max_{\alpha_i,\beta_i} L \\
&= \max_{\alpha_i,\beta_i} \min_{w,\xi_i} L \\
&= \max_{\alpha_i,\beta_i} \frac{1}{2\lambda} \left(\sum_{i=1}^{N} \beta_i y_i \phi(x_i) \right)^2 + \sum_{i=1}^{N} \beta_i \left(\xi_i + 1 - \xi_i - y_i w^T \phi(x_i) \right) \\
&= \max_{\alpha_i,\beta_i} \frac{1}{2\lambda} \left(\sum_{i=1}^{N} \beta_i y_i \phi(x_i) \right)^2 + \sum_{i=1}^{N} \beta_i - \frac{1}{\lambda} \sum_{i=1}^{N} \beta_i y_i \phi(x_i) \sum_{j=1}^{N} \beta_j y_j \phi(x_j) \\
&= \max_{\beta_i} -\frac{1}{2\lambda} \left[\sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j y_i y_j \phi(x_i)^T \phi(x_j) \right] + \sum_{i=1}^{N} \beta_i \\
&= \max_{\beta_i} -\frac{1}{2\lambda} \left[\sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j y_i y_j k(x_i, x_j) \right] + \sum_{i=1}^{N} \beta_i \\
s.t. \quad 0 \leq \beta_i \leq 1
\end{aligned}$$

In conclusion, the dual problem is

$$\begin{aligned}
\max_{\alpha_i,\beta_i} -\frac{1}{2\lambda} \left[\sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j y_i y_j k(x_i, x_j) \right] + \sum_{i=1}^{N} \beta_i \\
s.t. \quad 0 \leq \beta_i \leq 1
\end{aligned} \tag{10}$$

Accroding to KKT conditions:

$$\begin{aligned}
\hat{\beta}_i (y_i \hat{w}^T \phi(x_i) - 1 + \hat{\xi}_i) = 0 \\
\hat{w} = \frac{1}{\lambda} \sum_{i=1}^{N} \hat{\beta}_i y_i \phi(x_i)
\end{aligned}$$

where $\{\hat{\beta}_i\}$ are the optimal solution for Equation (10).

3.

$$\begin{aligned}
f(x) &= sign(\hat{w}^T \phi(x)) \\
&= sign\left(\frac{1}{\lambda} \sum_{i=1}^{N} \hat{\beta}_i y_i \phi(x_i)^T \phi(x) \right) \\
&= sign\left(\frac{1}{\lambda} \sum_{i=1}^{N} \hat{\beta}_i y_i k(x_i, x) \right)
\end{aligned} \tag{11}$$

where $\{\hat{\beta}_i\}$ are the optimal solution for Equation (10). Accroding to KKT conditions, we have either $\beta_i = 0$ or $\beta_i > 0, y_i \hat{w}^T \phi(x_i) = 1 - \hat{\xi}_i$, where $\hat{\xi}_i = 0$ indicates x_i is a margin support vector, whereas $\hat{\xi}_i > 0$ indicates x_i is an nonmargin support vector.

4 Boosting: from Weak to Strong

Problem 4

1.

Obviously,

$$\begin{aligned}\mathbb{I}\{y=-1\} &= \frac{1-y}{2}\\ \mathbb{I}\{y=1\} &= \frac{1+y}{2}\end{aligned}$$

because $\{x^{(i)}\}$ are in **descending** order, so for each threshold s, there is some $m_0(s) \in \{0,1,\cdots,m\}$ such that

$$\begin{aligned}\mathbb{I}\{\phi_{s,+}\left(x^{(i)}\right) \neq y^{(i)}\} &= \mathbb{I}\{y^{(i)}=1\}, \forall i > m_0(s)\\ \mathbb{I}\{\phi_{s,+}\left(x^{(i)}\right) \neq y^{(i)}\} &= \mathbb{I}\{y^{(i)}=-1\}, \forall i \leq m_0(s)\end{aligned}$$

thus

$$\begin{aligned}\sum_{i=1}^{m} p_i \mathbb{I}\{\phi_{s,+}\left(x^{(i)}\right) \neq y^{(i)}\} &= \sum_{i=1}^{m_0(s)} p_i \mathbb{I}\{y^{(i)}=-1\} + \sum_{i=m_0(s)+1}^{m} p_i \mathbb{I}\{y^{(i)}=1\}\\ &= \sum_{i=1}^{m_0(s)} p_i \frac{1-y^{(i)}}{2} + \sum_{i=m_0(s)+1}^{m} p_i \frac{1+y^{(i)}}{2}\\ &= \frac{1}{2} - \frac{1}{2}\left(\sum_{i=1}^{m_0(s)} y^{(i)} p_i - \sum_{i=m_0(s)+1}^{m} y^{(i)} p_i\right)\end{aligned} \tag{12}$$

Similarly, for each threshold s, there is some $m_0(s) \in \{0,1,\cdots,m\}$ such that

$$\begin{aligned}\mathbb{I}\{\phi_{s,-}\left(x^{(i)}\right) \neq y^{(i)}\} &= \mathbb{I}\{y^{(i)}=-1\}, \forall i > m_0(s)\\ \mathbb{I}\{\phi_{s,-}\left(x^{(i)}\right) \neq y^{(i)}\} &= \mathbb{I}\{y^{(i)}=1\}, \forall i \leq m_0(s)\end{aligned}$$

thus

$$\begin{aligned}\sum_{i=1}^{m} p_i \mathbb{I}\{\phi_{s,-}\left(x^{(i)}\right) \neq y^{(i)}\} &= \sum_{i=1}^{m_0(s)} p_i \mathbb{I}\{y^{(i)}=1\} + \sum_{i=m_0(s)+1}^{m} p_i \mathbb{I}\{y^{(i)}=-1\}\\ &= \sum_{i=1}^{m_0(s)} p_i \frac{1+y^{(i)}}{2} + \sum_{i=m_0(s)+1}^{m} p_i \frac{1-y^{(i)}}{2}\\ &= \frac{1}{2} - \frac{1}{2}\left(\sum_{i=m_0(s)+1}^{m} y^{(i)} p_i - \sum_{i=1}^{m_0(s)} y^{(i)} p_i\right)\end{aligned} \tag{13}$$

2.

We noted that

$$\begin{aligned}|f(m_0) - f(m_0+1)| &= \left|\sum_{i=1}^{m_0} y^{(i)} p_i - \sum_{i=m_0+1}^{m} y^{(i)} p_i - \sum_{i=1}^{m_0+1} y^{(i)} p_i + \sum_{i=m_0+2}^{m} y^{(i)} p_i\right|\\ &= |-2y^{m_0+1} p_{m_0+1}|\\ &= 2p_{m_0+1}\end{aligned}$$

because $\sum_{i=1}^m p_i = 1, p_i \geq 0, \forall i$, so $p_i \geq 1/m, \exists i$. Hence, we have

$$\max_{0 \leq m_0 \leq m-1} |f(m_0) - f(m_0+1)| \geq \frac{2}{m}$$

And we noted that

$$\begin{aligned}
\max_{m_0} |f(m_0)| &\geq \frac{1}{2} \max_{m_0} (|f(m_0)| + |f(m_0+1)|) \\
&\geq \frac{1}{2} \max_{m_0} |f(m_0) - f(m_0+1)| \\
&\geq \frac{1}{m} \\
&= 2\gamma
\end{aligned}$$

Hence

$$\max_{m_0} |f(m_0)| \geq 2\gamma \tag{14}$$

3.

Obviously,

$$\begin{aligned}
2\gamma &\leq \max_{m_0} |f(m_0)| \\
&= \max_{m_0} \left| \sum_{i=1}^{m_0} y^{(i)} p_i - \sum_{i=m_0+1}^{m} y^{(i)} p_i \right| \\
&\leq \left| \sum_{i=1}^{m_0} y^{(i)} p_i \right| + \left| \sum_{i=m_0+1}^{m} y^{(i)} p_i \right| \\
&\leq \sum_{i=1}^{m} p_i \\
&= 1
\end{aligned}$$

Hence, we have

$$\gamma \leq 1/2 \tag{15}$$

when all datas are classified correctly, $\gamma = 1/2$.

If there is only one stump, the error rate

$$J_1 \leq \frac{1}{2} - \gamma$$

thus

$$J_t \leq (1-4\gamma^2)^{\frac{1}{2}} J_{t-1} \leq \cdots \leq (1-4\gamma^2)^{\frac{t-1}{2}} J_1 \leq (1-4\gamma^2)^{\frac{t-1}{2}} (\frac{1}{2} - \gamma)$$

Let

$$J_t \leq (1-4\gamma^2)^{\frac{t-1}{2}} (\frac{1}{2} - \gamma) \leq \frac{1}{m}$$

we have

$$t \leq \frac{2 \ln \frac{2}{m(1-2\gamma)}}{\ln(1-4\gamma^2)} + 1$$

so the upper-bound $t_{upper-bound}$ is

$$t_{upper-bound} = \left\lfloor \frac{2 \ln \frac{2}{m(1-2\gamma)}}{\ln(1-4\gamma^2)} + 1 \right\rfloor \tag{16}$$