CS229 Problemset 02

KOH-

April 2025

Problem 01

a)

The **Logistic Regression** converges in Dataset A, but doesn't converge in Dataset B.

b)

After plot each dataset, I found that Dataset A is **not linearly separable**, but Dataset B is **linearly separable**.

The difference is that:

The functional margin for Dataset A:

$$\hat{\gamma_A} = \min(y^{(i)}(\omega_A^T x^{(i)} + b))$$

Since DatasetA is not linearly seperable, therefore

$$\max(\hat{\gamma_A}) < 0$$

However, for a linearly seperable Dataset B:

$$\hat{\gamma_B} = \min(y^{(i)}(\omega_B^T x^{(i)} + b))$$
$$max(\hat{\gamma_B}) > 0$$

For the linear regression, we need to maximize the likelihood:

$$L(\theta) = \prod_{i=1}^{m} P(y|x,\theta) = \prod_{i=1}^{m} h_\theta(x^{(i)})$$
$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{m}(-\log(1 + e^{-y^{(i)}\omega^T x^{(i)}}))$$

So we need to minimize:

$$\sum_{i=1}^{m} \log(1 + e^{-y^{(i)}\omega^T x^{(i)}})$$

Since we do not confine $||\omega||$ and for dataset B:

$$\min(y^{(i)}(\omega^T x^{(i)})) > 0$$

So we can replace ω with 2ω, it can make $\ell(\theta)$ smaller, so it will not converge for dataset B.

But for Dataset A, if we keep making ω larger, those $y^{(i)}(\omega_A^T x^{(i)} + b) < 0$ will make $\ell(\theta)$ much bigger, but we want to minimize it, so it will converge.

c)

For each of these possible modifications, state whether or not it would lead to the provided training algorithm converging on datasets such as B. Justify your answers

i)

It does not work.

If the learning rate is too small, it will converge too early.

If it is too large, it will not solve the problem that ω can increase to inf.

ii)
No, I think all the changes on learning rate will not solve the core problem that the ω can converge to inf.
iii.
No, the linearly seperable dataset will still be linearly seperable, which also allows ω to converge to inf.
iv.
Yes, it will definitely help to avoid $||\omega|| \to +\infty$
v.
I think Yes, the method will make the original dataset become not linearly seperable.

d)

SVM with hinge loss are not vulnerable to datasets like B.

The hinge loss is :

$$\ell(\theta) = \max(0, 1 - y^{(i)}\hat{y})$$

This means if we get every functional margin more than 1, the progress will stop.

Due to the dataset is linearly seperable, there exists a ω such that $\forall i\ y^{(i)}(\omega x^{(i)} + b) > 0$. therefore we can rescale ω to let every $|\hat{y}| \geq 1$, then the whole progress will stop.

Problem 02

a)

The likelihood function of the linear regression is:

$$\ell(\theta) = \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))$$

After training, the partial derivative will be 0, which means:

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) \cdot x_j^{(i)} = 0$$

Consider $j = 0$, and for all i, $x_j^{(i)} = 0$, then

$$\sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) = 0 \to \sum_{i=1}^{m} y^{(i)} = \sum_{i=1}^{m} h_\theta(x^{(i)}) = \sum_{i=1}^{m} P(y = 1|x, \theta)$$

When $(a, b) = (0, 1)$, $I_{a,b} = \{1.....m\}$, so:

$$\sum_{i \in I_{a,b}} P(y^{(i)} = 1|x^{(i)}; \theta) = \sum_{i \in I_{a,b}} \mathbb{I}\{y^{(i)=1}\}$$

b)

No, perfectly calibrated does not imply the model achieves perfect accuracy. The converse is not necessarily true either.

In our prediction, we let every $\hat{y^{(i)}} = 1$ If $h_\theta(x^{(i)}) \geq 0.5$

For example, many examples' predicted probability is 0.7, but in our prediction they are all $\hat{y^{(i)}} = 1$, but there are 30% of them with $y^{(i)} = 0$. Therefore, perfectly calibrated does not imply the model achieves perfect accuracy.

The converse is not necessarily true either.

We can explain in mathematics, consider $(a, b) = (0.5, 1)$. For a perfectly accurate model, obviously:

$$\sum_{i \in I_{a,b}} P(y^{(i)}|x^{(i)}; \theta) < \sum_{i \in I_{a,b}} \mathbb{I}\{y^{(i)} = 1\}$$

But for a perfectly calibrated model: for any $(a, b) \subset [0, 1]$

$$\sum_{i \in I_{a,b}} P(y^{(i)}|x^{(i)}; \theta) = \sum_{i \in I_{a,b}} \mathbb{I}\{y^{(i)} = 1\}$$

So this statement and its reverse are not true.

c)

The L_2 regularization in the logistic regression is:

$$J(\theta) = -\ell(\theta) + \lambda||\theta||_2^2$$

After training, the gradient will be 0.

$$\frac{\partial J(\theta)}{\partial \theta_j} = \sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})x_j^{(i)} + 2\lambda\theta_j$$

Let $j = 0$

$$\frac{\partial J(\theta)}{\partial \theta_0} = \sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)}) + 2\lambda\theta_0 = 0$$

So:

$$\sum_{i=1}^{m} h_\theta(x^{(i)}) + \lambda\theta_0 = \sum_{i=1}^{m} y^{(i)}$$

But $\theta_0 = 0$ not always holds, so the original calibration over $(a, b) = (0, 1)$ will be changed.

Problem 03

a)

$$\begin{aligned}
\theta_{MAP} &= \arg\max_\theta p(\theta|x, y) \\
p(\theta|x, y) &= \frac{p(\theta, x, y)}{p(x, y)} = \frac{p(y|x, \theta)p(x, \theta)}{p(x, y)} \\
&= \frac{p(y|x, \theta)p(\theta|x)p(x)}{p(x, y)} = p(y|x, \theta)p(\theta)\frac{p(x)}{p(x, y)}
\end{aligned}$$

$\frac{p(x)}{p(x,y)}$ is a constant.
So

$$\theta_{MAP} = \arg\max_\theta p(\theta|x, y) = \arg\max_\theta p(y|x, \theta)p(\theta)$$

b)

$$\begin{aligned}
\theta_{MAP} &= \arg\max_\theta p(y|x, \theta)p(\theta) \\
&= \arg\min_\theta -p(y|x, \theta)p(\theta) \\
&= \arg\min_\theta -\log p(y|x, \theta) - \log p(\theta)
\end{aligned}$$

Because $\theta \sim N(0, \eta^2 I)$

$$p(\theta) = \frac{1}{(2\pi)^{n/2}\eta^n}\exp\{-\frac{1}{2\eta^2}\theta^T\theta\}$$

$$\log p(\theta) = -\frac{n}{2}\log(2\pi) - n\log\eta - \frac{1}{2\eta^2}||\theta||_2^2$$

Because $\log p(\theta) = -\frac{n}{2}\log(2\pi) - n\log\eta$ is a constant, consider $\lambda = \frac{1}{2\eta^2}$

$$\theta_{MAP} = \arg\min_\theta -\log p(y|x, \theta) + \lambda||\theta||_2^2$$

c)

$$
\begin{aligned}
&y^{(i)} = \theta^T x + \epsilon, \epsilon \sim N(0, \sigma^2) \\
&y^{(i)} \sim N(\theta^T x^{(i)}, \sigma^2) \\
&p(y^{(i)}|x^{(i)}, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\} \\
&\log p(\vec{y}|X, \theta) = \sum_{i=1}^{m} (\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}) \\
&\theta_{MAP} = \arg\min_{\theta} - \log p(\vec{y}|X, \theta) + \frac{1}{2\eta^2}||\theta||_2^2
\end{aligned}
$$

$$
\begin{aligned}
J(\theta) &= -\log p(\vec{y}|X, \theta) + \frac{1}{2\eta^2}||\theta||_2^2 \\
&= -\sum_{i=1}^{m} (\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}) + \frac{1}{2\eta^2}||\theta||_2^2
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_j} &= \sum_{i=1}^{m} \frac{(\theta^T x^{(i)} - y^{(i)}) x_j^{(i)}}{\sigma^2} + \frac{\theta_j}{\eta^2} \\
\nabla_\theta J(\theta) &= \frac{1}{\sigma^2} X^T (X\theta - \vec{y}) + \frac{1}{\eta^2}\theta
\end{aligned}
$$

We want to minimize $J(\theta)$, so $\nabla_\theta J(\theta) = 0$
So

$$\theta_{MAP} = (X^T X + \frac{\sigma^2}{\eta^2} I)^{-1} X^T \vec{y}$$

d)

$$\theta_{MAP} = \arg\max_{\theta} p(y|x, \theta) p(\theta)$$

$$
\begin{aligned}
&y^{(i)}|x, \theta \sim N(\theta^T x^{(i)}, \sigma^2) \\
&p(\vec{y}|X, \theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(\theta^T x^{(i)} - y^{(i)})^2}{2\sigma^2}\} \\
&p(\theta) = \frac{1}{(2b)^n} \exp\{\sum_{i=1}^{n} -\frac{|\theta_i|}{b}\} = \frac{1}{(2b)^n} \exp\{-\frac{1}{b}||\theta||_1\} \\
&\theta_{MAP} = \arg\max_{\theta} p(\vec{y}|X, \theta) p(\theta) = \arg\min_{\theta} - \log p(\vec{y}|X, \theta) - \log p(\theta) \\
&\theta_{MAP} = \arg\min_{\theta} \frac{1}{2\sigma^2}||X\theta - \vec{y}||_2^2 + \frac{1}{b}||\theta||_1
\end{aligned}
$$

So

$$\theta_{MAP} = \arg\min_{\theta} ||X\theta - \vec{y}||_2^2 + \frac{2\sigma^2}{b}||\theta||_1, \gamma = \frac{2\sigma^2}{b}$$

Problem 04

a)

Yes.

K_1, K_2 are kernels over $\mathbb{R}^n \times \mathbb{R}^n$. So K_1, K_2 are PSD.

So:

$z^T K z = z^T (K_1 + K_2) z = z^T K_1 z + z^T K_2 z \geq 0$

So K is PSD, so K is a kernel.

b)

Not always.

$K_1 - K_2$ may not be PSD.

c)

Yes.

$a \in \mathbb{R}^+$

so $z^T K z = z^T a K_1 z = a z^T K z \geq 0$

d)

No, $z^T K z \leq 0$

e)

$$
\begin{aligned}
z^T K z &= \sum_i \sum_j z_i K_{ij} z_j \\
&= \sum_i \sum_j z_i K_1(x^{(i)}, x^{(j)}) K_2(x^{(i)}, x^{(j)}) z_j \\
&= \sum_i \sum_j z_i \phi_1(x^{(i)})^T \phi_1(x^{(j)}) \cdot \phi_2(x^{(i)})^T \phi_2(x^{(j)}) z_j \\
&= \sum_i \sum_j z_i \left(\sum_a \phi_{1a}(x^{(i)}) \phi_{1a}(x^{(j)}) \right) \left(\sum_b \phi_{2b}(x^{(i)}) \phi_{2b}(x^{(j)}) \right) z_j \\
&= \sum_a \sum_b \sum_i \sum_j z_i \phi_{1a}(x^{(i)}) \phi_{1a}(x^{(j)}) \phi_{2b}(x^{(i)}) \phi_{2b}(x^{(j)}) z_j \\
&= \sum_a \sum_b \left(\sum_i z_i \phi_{1a}(x^{(i)}) \phi_{2b}(x^{(i)}) \right)^2 \geq 0
\end{aligned}
$$

f)

$$K_{ij} = f(x^{(i)}) \cdot f(x(j)) \tag{1}$$

$$z^T K z = \sum_i \sum_j z_i K_{ij} z_j = \sum_i \sum_j z_i f(x^{(i)}) f(x^{(j)}) z_j = \left(\sum_i z_i f(x^{(i)}) \right)^2 \geq 0 \tag{2}$$

g)

It's the same as the definition.

h)

$$\begin{aligned} K(x,z) &= p\left(K_1(x,z)\right) \\ &= \sum_i c_i \left(K_1(x,z)\right)^i \end{aligned}$$

According to (e) and (c).Because $c_i > 0$ $c_i \left(K_1(x,z)\right)^i$ is a valid kernel.

At last (a) suggests:

$$K(x,z) = p\left(K_1(x,z)\right) = \sum_i c_i \left(K_1(x,z)\right)^i$$

is a valid kernel.

Problem 05

a)

i.

$$\theta^{(i)} = \sum_{j=1}^{i} \alpha_j \phi(x^{(j)})$$

so $\theta^{(0)} = \sum_{j=1}^{0} \alpha_j \phi(x^{(j)}) = \vec{0}$

ii.

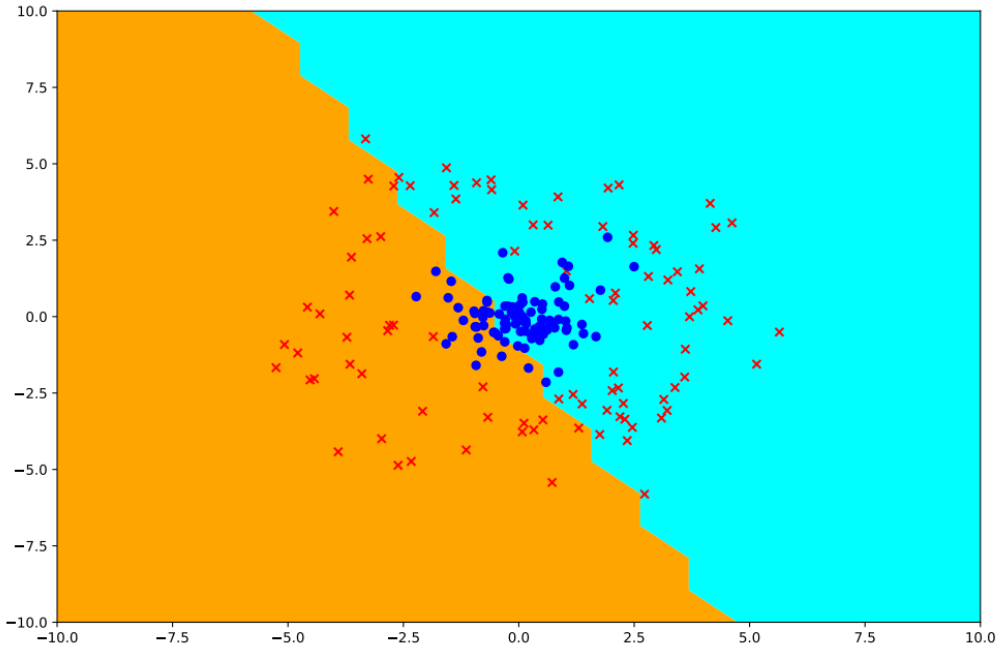
$$\begin{aligned} h_{\theta^{(i)}}(x^{(i+1)}) &= sign\left(g({\theta^{(i)}}^T \phi(x^{(i+1)}))\right) \\ &= sign\left(g\left((\sum_{j=1}^{i} \alpha_j \phi(x^{(j)})) \cdot \phi(x^{(i+1)})\right)\right) \\ &= sign\left(g\left(\sum_{j=1}^{i} \alpha_j \phi(x^{(j)}) \cdot \phi(x^{(i+1)})\right)\right) \\ &= sign\left(g\left(\sum_{j=1}^{i} \alpha_j K(x^{(j)}, x^{(i+1)})\right)\right) \end{aligned}$$

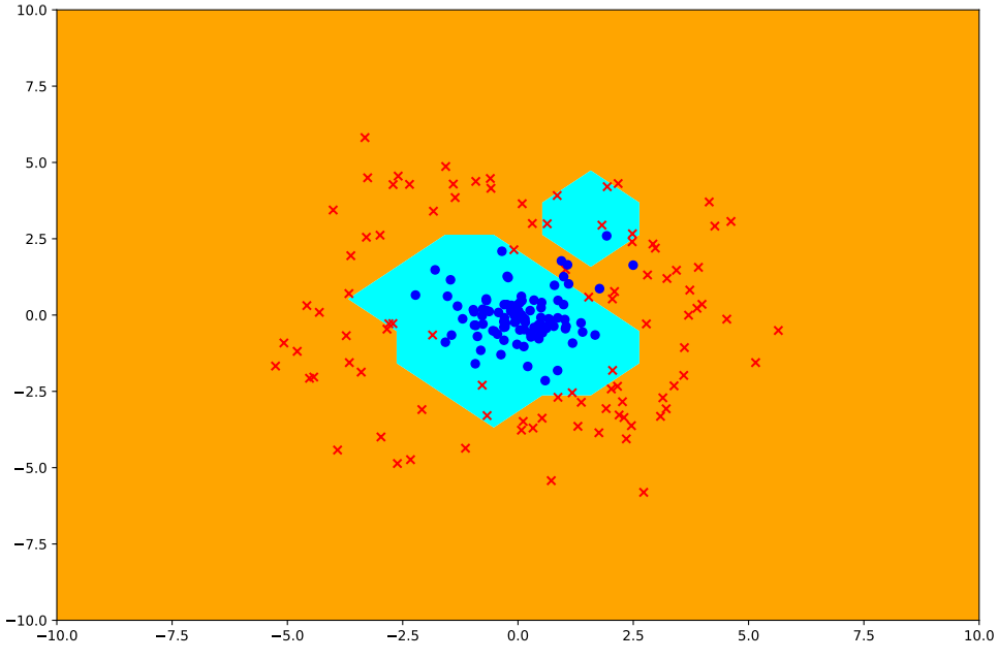
iii.

As I mentioned above, we don't need to know what $\theta^{(i)}$ is. We only need to know the coefficients' value.

so $\alpha_j = \alpha(y^{(i+1)} - h_{\theta^{(i)}}(x^{(i+1)})) = \alpha(y^{(i+1)} - sign\left(g({\theta^{(i)}}^T \phi(x^{(i+1)}))\right))$

b)





c)

dot kernel performs extremely poorly in classifying the points.

Because the test data in not linear separable,. The dot kernels is the basic perceptron with $\phi(x) = x$ which is used to classify the linear separable data.

Problem 06

b)

$$
\begin{aligned}
p(y=1|x) &= \frac{\prod_{i=1}^{m} p(x_i|y=1)p(y=1)}{\prod_{i=1}^{m} p(x_i|y=1)p(y=1) + \prod_{i=1}^{m} p(x_i|y=0)p(y=0)} \\
&= \frac{1}{1 + \frac{\prod_{i=1}^{m} p(x_i|y=0)p(y=0)}{\prod_{i=1}^{m} p(x_i|y=1)p(y=1)}} \\
p(y=1|x) > 0.5 &\rightarrow \frac{\prod_{i=1}^{m} p(x_i|y=0)p(y=0)}{\prod_{i=1}^{m} p(x_i|y=1)p(y=1)} > 1 \\
\prod_{i=1}^{m} p(x_i|y=0)p(y=0) &> \prod_{i=1}^{m} p(x_i|y=1)p(y=1) \\
\sum_{i=1}^{m} \log p(x_i|y=0) + \log p(y=0) &> \sum_{i=1}^{m} \log p(x_i|y=1) + \log p(y=1)
\end{aligned}
$$

c)

The top 5 indicative words for Naive Bayes are: ['claim', 'won', 'prize', 'tone', 'urgent!']

d)

The optimal SVM radius was 0.1.

The SVM model had an accuracy of 0.9695340501792115 on the testing set.