DDA3020 Machine Learning Spring 2023
Produced by TA Teams

Assignment 2

1 Problem 1

The first term of $J(\mathbf{w})$'s Hessian is positive definite(8.7), the second term's Hessian is positive definite as well ($\lambda > 0$). Therefore this function has a positive definite Hessian, it has a global optimum. The form of posterior distribution takes:

$$p(\mathbf{w} \mid D) \propto p(D \mid \mathbf{w})p(\mathbf{w})$$

$$p(\mathbf{w}) = N\left(\mathbf{w} \mid \mathbf{0}, \sigma^{-2}\mathbf{I}\right)$$

$$NLL(\mathbf{w}) = -\log p(\mathbf{w} \mid D) = -\log p(D \mid \mathbf{w}) + \frac{1}{2\sigma^2}\mathbf{w}^T\mathbf{w} + c$$

Therefore,

$$\lambda = \frac{1}{2\sigma^2}$$

- **F** To show this, we could show that $\mathbf{H}_-$ is positive definite and therefore the regularized loss is strictly convex. Since

  $$\mathbf{H} = \sum_i \left(\mu_i \left(1 - \mu_i\right) \mathbf{x}_i \mathbf{x}_i^T\right) + 2\lambda$$

  is positive definite, $J(\mathbf{w})$ has a unique global optimum.

- **F** ℓ_2 regularization tends to penalize larger weights more heavily, but this does not need to imply sparsity of the global optimum.

- **T** If $\lambda = 0$, the optimum can be a step function, similar to a logistic function where $\|\mathrm{w}\| \to \infty$

- **F** If λ increases, bias for both train set and test set increase, therefore log likelihood decreases.

- **F** If λ increases, bias for both train set and test set increase, therefore log likelihood decreases.

2 Problem 2

2.1 Consider a potential solution to the above problem with some $\xi < 0$. Then the constraint $y^{(I)}(w^T x^{(i)} + b) \geq 1 - \xi_i$i would also be satisfied for $\xi_i = 0$, and the objective function would be lower, proving that this could not be an optimal solution.
2.2

Answer:

$$\mathcal{L}(w, b, \xi, \alpha) = \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^{m} \xi_i^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i],$$

where $\alpha_i \geq 0$ for $i = 1, \ldots, m$.

2.3 $w = \sum_{i=1}^{m} a_i y^{(i)} x^{(i)}$, $0 = \sum_{i=1}^{m} a_i y^{(i)}$, $C\xi_i = a_i$
2.4

Answer: The objective function for the dual is

$$\begin{aligned}
W(\alpha) &= \min_{w,b,\xi} \mathcal{L}(w, b, \xi, \alpha) \\
&= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i y^{(i)} x^{(i)})^T (\alpha_j y^{(j)} x^{(j)}) + \frac{1}{2} \sum_{i=1}^{m} \frac{\alpha_i}{\xi_i} \xi_i^2 \\
&\quad - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} \left(\left(\sum_{j=1}^{m} \alpha_j y^{(j)} x^{(j)} \right)^T x^{(i)} + b \right) - 1 + \xi_i \right] \\
&= -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} + \frac{1}{2} \sum_{i=1}^{m} \alpha_i \xi_i \\
&\quad - \left(\sum_{i=1}^{m} \alpha_i y^{(i)} \right) b + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i \xi_i \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - \frac{1}{2} \sum_{i=1}^{m} \alpha_i \xi_i \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - \frac{1}{2} \sum_{i=1}^{m} \frac{\alpha_i^2}{C}.
\end{aligned}$$

Then the dual formulation of our problem is

$$\begin{array}{ll}
\max_{\alpha} & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} - \frac{1}{2} \sum_{i=1}^{m} \frac{\alpha_i^2}{C} \\
\text{s.t.} & \alpha_i \geq 0, \ i = 1, \ldots, m \\
& \sum_{i=1}^{m} \alpha_i y^{(i)} = 0
\end{array}.$$

3 Problem 3

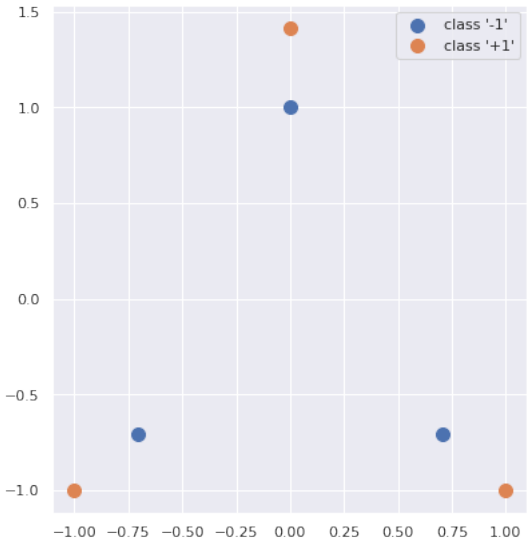


Figure 1: Visualization of Data Set

1. Since the data is not linearly separable, we cannot find an SVM classifier without slack variable for this dataset.
   (It also makes sense to say that it's possible to find one when we are able to add a suitable nonlinear kernel)

2. After expanding the original feature vector, we get the new labels:

$$\text{Class -1: } \begin{bmatrix} (0 & 1) \\ (\frac{1}{2} & \frac{1}{2}) \\ (\frac{1}{2} & \frac{1}{2}) \end{bmatrix} \qquad \text{Class } +1: \begin{bmatrix} (0 & 2) \\ (1 & 1) \\ (1 & 1) \end{bmatrix}$$

   By plotting the data, we can easily find that all the data are support vectors for the SVM. And the margin of the SVM is $\frac{\sqrt{2}}{4}$ (Half of the distance between the two black lines.

   Since $w_1x_1 + w_2x_2 + b = 0$ when $x_1 + x_2 = \frac{3}{2}$:

$$w_1x_1 + (-x_1 + \frac{3}{2})w_2 + b = 0$$
$$(w_1 - w_2)x_1 + (b + \frac{3}{2}w_2) = 0$$

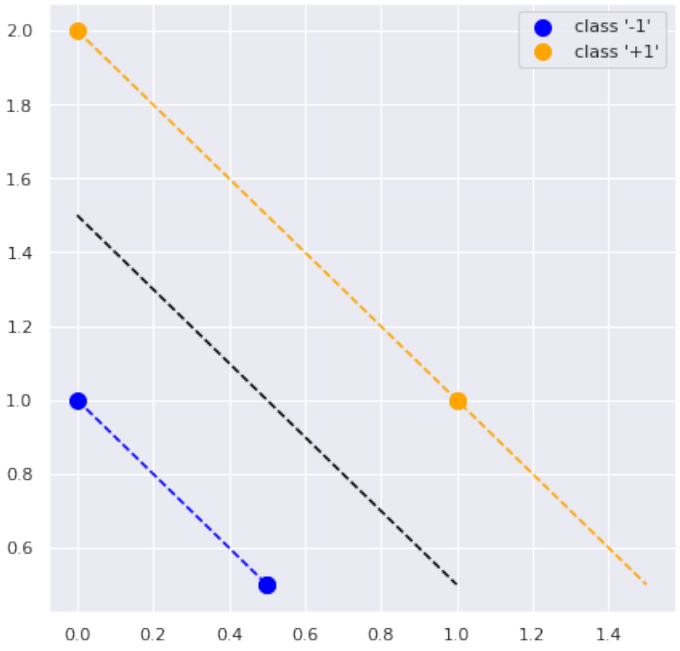


Figure 2: Visualization of Expanded Data Set

Then, we get $w_1 = w_2$ and $b = -\frac{3}{2}w_2$.

Thus, $margin = \frac{1}{||\mathbf{w}||} = \frac{1}{\sqrt{w_1^2+w_2^2}} = \frac{\sqrt{2}}{4}$. We get $w_1 = w_2 = 2$ and $b = -3$.

Thus the decision boundary for the SVM is $f(\phi(\mathbf{x})) = (2,2)\phi(\mathbf{x}) - 3$

We will predict class +1 when $f(\phi(\mathbf{x})) > 0$ and class -1 otherwise.

Predicting for $(-\frac{1}{\sqrt{2}}, \sqrt{2})$:

$$\begin{aligned} f(\phi((-\frac{1}{\sqrt{2}}, \sqrt{2}))) &= (2,2)\phi((-\frac{1}{\sqrt{2}}, \sqrt{2})) - 3 \\ &= (2,2)(\frac{1}{2}, 2)^T - 3 \\ &= 1 + 4 - 3 = 2 > 0 \end{aligned}$$

Thus $(-\frac{1}{\sqrt{2}}, \sqrt{2})$ belongs to class +1.

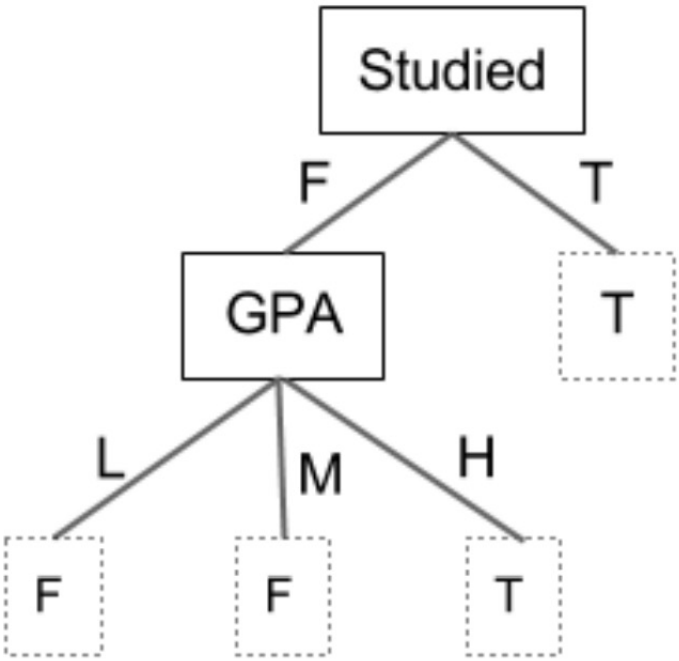
4 Problem 4

1.
$$\begin{aligned} H(\text{ Passed }) &= -\left(\frac{2}{6}\log_2\frac{2}{6} + \frac{4}{6}\log_2\frac{4}{6}\right) \\ H(\text{ Passed }) &= -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) \\ H(\text{ Passed }) &= \log_2 3 - \frac{2}{3} \approx 0.92 \end{aligned} \tag{1}$$

2.
$$\begin{aligned} H(\text{ Passed } \mid GPA) &= -\frac{1}{3}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{1}{3}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) - \frac{1}{3}\left(1\log_2 1\right) \\ H(\text{ Passed } \mid GPA) &= \frac{1}{3}(1) + \frac{1}{3}(1) + \frac{1}{3}(0) \\ H(\text{ Passed } \mid GPA) &= \frac{2}{3} \approx 0.66 \end{aligned} \tag{2}$$

3.
$$\begin{aligned} H(\text{ Passed } \mid \text{ Studied }) &= -\frac{1}{2}\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{1}{2}\left(1\log_2 1\right) \\ H(\text{ Passed } \mid \text{ Studied }) &= \frac{1}{2}\left(\log_2 3 - \frac{2}{3}\right) \\ H(\text{ Passed } \mid \text{ Studied }) &= \frac{1}{2}\log_2 3 - \frac{1}{3} \approx 0.46 \end{aligned}$$

4. We want to split first on the variable which maximizes the information gain $H(\text{ Passed }) - H(\text{ Passed } \mid A)$. This is equivalent to minimizing $H(\text{ Passed } \mid A)$, so we should split on "Studied?" first.


Studied
F
T
GPA
T
L
M
H
F
F
T