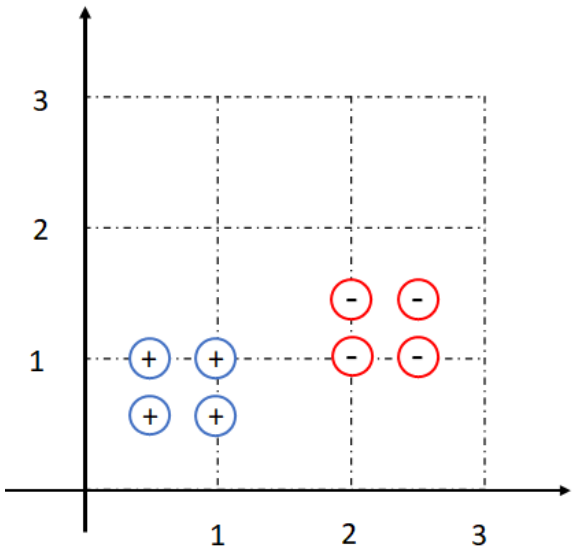
Submitted by: Oluwasegun Somefun, somefuno@oregonstate.edu

AI534 — Written Homework Assignment 3 (40 pts) — Due Nov 12th, 2021

1. Apply linear SVM without soft margin to the following problem.



a. (2pts) Please mark out the support vectors, the decision boundary ($\mathbf{w}^T\mathbf{x}+b=0$) and $\mathbf{w}^T\mathbf{x}+b=1$ and $\mathbf{w}^T\mathbf{x}+b=-1$. Note that you don't need to solve the optimization problem for this, just eyeball the solution.

The support vectors are circled in black ink, and the decision boundaries are clearly marked out in Figure 1 The support vectors, that can be observed for $\mathbf{x}=(\mathbf{x}_1,\mathbf{x}_2)$ are points on the decision

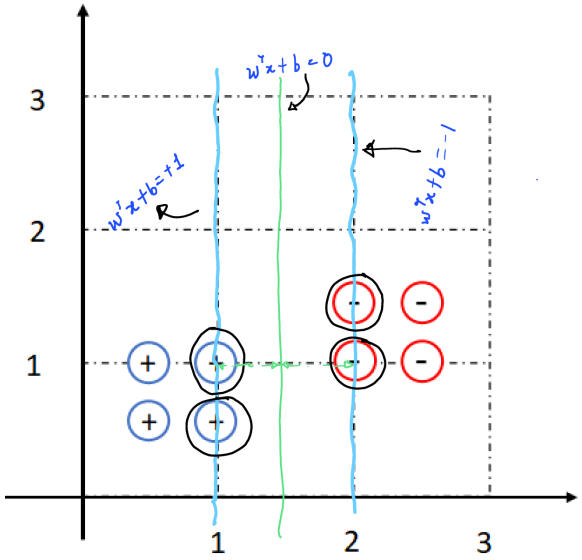


Figure 1: Marked SVM Decision Boundary

boundary, For the $y = +1$ class, they are potentially $\mathbf{x} = (1,1), (1,0.5)$, and for the $y = -1$ class, we have $\mathbf{x} = (2,1), (2,1.5)$

The exact support vectors with the smallest margin are $\mathbf{x} = (1,1), (2,1)$.

The decision boundary $\mathbf{w}^T\mathbf{x} + b = 0$ is the vertical straight-line passing through $\mathbf{x} = (1.5, 0)$.

The decision boundary $\mathbf{w}^T\mathbf{x} + b = +1$ is the vertical straight-line passing through $\mathbf{x} = (1, 0)$.

The decision boundary $\mathbf{w}^T\mathbf{x} + b = -1$ is the vertical straight-line passing through $\mathbf{x} = (2, 0)$.

b. (6 pts) Please solve for $\mathbf{w}$ and b based on the support vectors you identified in (a).

The exact decision boundary line for the half-space is $\mathbf{w}^T\mathbf{x} + b = 0$, which passes through $\mathbf{x} = (1.5, 0)$. Solving, we have:

$$\begin{pmatrix} b \\ \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} = \begin{pmatrix} b \\ \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} \begin{pmatrix} 1 \\ 1.5 \\ 0 \end{pmatrix} = 0$$

$$b = -1.5\mathbf{w}_1$$

For each halfspace, supported by the vectors $\mathbf{x} = (1,1), (2,1)$ we have:

$$\begin{pmatrix} -1.5\mathbf{w}_1 \\ \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 1 \quad \text{and} \quad \begin{pmatrix} -1.5\mathbf{w}_1 \\ \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = -1$$

Solving the resulting simultaenous equation we obtain $\mathbf{w}_1 = -2, \mathbf{w}_2 = 0, b = 3$ and hence

$$\mathbf{w} = (-2, 0) \quad \text{and} \quad b = 3$$

2. L_2 SVM

Given a set of training examples $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $y_i \in \{1, -1\}$ for all i. The following is the primal formulation of L_2 SVM, a variant of the standard SVM obtained by squaring the hinge loss:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \ & \mathbf{w}^T\mathbf{w} + \lambda \sum_{i=1}^{N} \xi_i^2 \\ \text{s.t. } & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \ i \in \{1, \cdots, N\} \\ & \xi_i \geq 0, \ i \in \{1, \cdots, N\} \end{aligned}$$

a. (5pts) Show that removing the second constraint $\xi_i \geq 0$ will not change the solution to the problem. In other words, let $(\mathbf{w}^*, b^*, \xi^*)$ be the optimal solution to the problem without this set of constraints, show that $\xi_i^* \geq 0, \forall i \in \{1, \cdots, N\}$. (Hint: use proof by contradiction.)

Proof by Contradiction. Assume for some $i \ \exists \ \xi_i^* < 0$, then the first constraint is satisfied and reduces to: $y_i(\mathbf{w}^{*T}\mathbf{x}_i + b) \geq 1$ which means there is a point outside the margin that does not contribute to the loss. If instead $\xi_i^* = 0$, the first constraint still satisfied reduces to: $y_i(\mathbf{w}^{*T}\mathbf{x}_i + b) \geq 1$ which means there is a point on the margin that does not contribute to the loss. Since these does not contribute to the loss, they cannot be the optimal solution. Therefore, explicitly stating the second constraint is redundant, removing it will not change the solution to the optimization problem, beacuse to contribute to the loss and obtain an optimal solution, the second constraint or condition: $\xi_i^* \geq 0 \ \forall i$ must hold, since, by definition $\xi_i^* = \max(0, 1 - y_i(\mathbf{w}^{*T}\mathbf{x} + b)) \geq 0$ □

b. (2 pts) After removing the second set of constraints, we have a simpler problem with only one set of constraints. Now provide the lagrangian of this new problem.

The Lagrangian L of this problem, with the corresponding Lagrange multipliers α_i can be provided as follows

$$L = \min_{\mathbf{w},b,\xi} \ \mathbf{w}^T\mathbf{w} + \lambda \sum_{i=1}^{N} \xi_i^2 + \sum_{i=1}^{N} \alpha_i (1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

c. (5pts) Derive the dual of this problem. How is it different from the standard SVM with hinge loss? Which formulation is more sensitive to outliers?

Let us start with differentiating L for parameters $\mathbf{w}, b, \xi_i$

$$\nabla_{\mathbf{w}} L = 2\mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = 0$$

$$\nabla_b L = \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\nabla_{\xi_i} L = 2\lambda\xi_i - \alpha_i = 0$$

We obtain:

$$\mathbf{w} = \frac{1}{2} \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i, \quad \xi_i = \frac{1}{2\lambda}\alpha_i$$

Substituting, these two parameters to be minimized back into expression L, we obtain the dual form

$$L_d = \max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \frac{1}{4\lambda} \sum_{i=1}^{N} \alpha_i^2$$
$$\text{s.t. } \alpha_i \geq 0 \quad \forall i$$

equivalently

$$L_d = \max_{\boldsymbol{\alpha}} \quad \boldsymbol{\alpha}^T\mathbf{1} - \frac{1}{2}(\boldsymbol{\alpha}\mathbf{y}\mathbf{x})^T(\mathbf{x}\boldsymbol{\alpha}\mathbf{y}) - \frac{1}{4\lambda}\boldsymbol{\alpha}^T\boldsymbol{\alpha}$$
$$\text{s.t. } \forall i \text{ in } \boldsymbol{\alpha}, \quad \alpha_i \geq 0$$

To compare, with the above, given below is the dual of the standard SVM with soft margin.

$$L_d = \max_{\boldsymbol{\alpha}} \quad \boldsymbol{\alpha}^T\mathbf{1} - \frac{1}{2}(\boldsymbol{\alpha}\mathbf{y}\mathbf{x})^T(\mathbf{x}\boldsymbol{\alpha}\mathbf{y})$$
$$\text{s.t. } \forall i \text{ in } \boldsymbol{\alpha}, \quad 0 \leq \alpha_i \leq \lambda$$

The dual of this variant of the SVM with soft margin has an extra $\mathcal{L}_2$ norm term on $\boldsymbol{\alpha}$ compared to the dual of the standard SVM with soft margin problem which is not squared.

Also, this formulation is not constrained by λ, compared to the dual of the standard SVM with soft margin formulation

Hence, intuitively, using the noted differences above, this variant of the SVM with soft-margin formulation may be more sensitive to outliers.

3. (Naive Bayes Classifier) Consider the following training set:

A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

(a) (5 pts) Learn a Naive Bayes classifier by estimating all necessary probabilities (there should be 7 probabilities in total).

The generative process is as follows

1. Learn the prior probability: for $j = 1, ..., k$, where $k = 2$

$$P(y = y_1 = 0) = 0.5, \quad P(y = y_2 = 1) = 0.5$$

2. Learn the probability distribution on $\mathbf{x}$ conditionally independent on each y_j for $i = 1, ..d$, where $d = 3$ in the set $\mathbf{x}_1 = A, \mathbf{x}_2 = B, \mathbf{x}_3 = C$

$$P(\mathbf{x}_1 = 1 \,|\, y_1) = \frac{1}{3}, \quad P(\mathbf{x}_1 = 1 \,|\, y_2) = \frac{2}{3}$$

$$P(\mathbf{x}_2 = 0 \,|\, y_1) = \frac{1}{3}, \quad P(\mathbf{x}_2 = 0 \,|\, y_2) = \frac{1}{3}$$

$$P(\mathbf{x}_3 = 1 \,|\, y_1) = \frac{1}{3}, \quad P(\mathbf{x}_3 = 1 \,|\, y_2) = \frac{2}{3}$$

(b) (5 pts) Compute the probability $P(y = 1|A = 1, B = 0, C = 0)$.

3. Compute $P(y \,|\, \mathbf{x} = (1, 0, 0))$ using the Naive Bayes formula based on conditional independence, we obtain

$$P(y = 1 \,|\, \mathbf{x} = (1, 0, 0)) = \frac{(\frac{2}{3}\frac{1}{3}\frac{2}{3})\frac{1}{2}}{\left[(\frac{2}{3}\frac{1}{3}\frac{2}{3})\frac{1}{2} + (\frac{1}{3}\frac{1}{3}\frac{1}{3})\frac{1}{2}\right]} = \frac{\frac{4}{54}}{\frac{5}{54}} = \frac{4}{5}$$

Similarly,

$$P(y = 0 \,|\, \mathbf{x} = (1, 0, 0)) = 1 - \frac{4}{5} = \frac{1}{5}$$

4. Since $P(y = 1 \,|\, \mathbf{x} = (1, 0, 0)) > P(y = 0 \,|\, \mathbf{x} = (1, 0, 0))$, then we can predict $P(y = 1 \,|\, \mathbf{x} = (1, 0, 0))$

(c) (2 pts) Suppose we know that A, B and C are independent random variables, can we say that the Naive Bayes assumption is valid? (Note that the particular data set is irrelevant for this question). If your answer is yes, please explain why; if you answer is no please give an counter example.

No! Suppose we know that A, B and C are independent random variables, then the Naive Bayes assumption is not valid? This is because indpendence does not imply conditional independence, which is assumed by the Naive Bayes classifier.

As a counter-example, see that for independence $P(x = (A, B, C)) = P(A)P(B)(C)$ while for conditional independence $P(x = (A, B, C)|y) = P(A|y)P(B|y)(C|y)$

4. (Naive Bayes learns linear decision boundary.) (8 pts) Consider a naive Bayes binary classifier with a set of binary features $x_1, x_2, ..., x_d$. Show that the Naive Bayes classifier learns a linear decision boundary $w_0 + w_1x_1 + w_2x_2 + ... + w_dx_d = 0$. Express the weights using the Naive Bayes parameters. Hint: consider the decision rule of predicting $y = 1$ if $P(y = 1|\mathbf{x}) > P(y = 1|\mathbf{x})$. This is equivalent to having a decision boundary defined by $\log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = 0$.

Proof: Naive-Bayes lerns a linear decision boundary. For $\mathbf{x}_i \in \mathbb{R}^d$, a linear decision boundary is given by a finite affine or linear combination: $\mathbf{w}^T\mathbf{x} + w_0 = 0$, where $\mathbf{w} \in \mathbb{R}^d$.

We start with the following assumptions:

(a) y is binary output with a bernoulli distribution with a prior $P(y = 1)$

(b) $\mathbf{x}_i$ is an input with binary features of a bernoulli distribution ${\theta_i}^{\mathbf{x}_i}(1 - \theta_i)^{1-\mathbf{x}_i}$

(c) the Naive bayes assumptions of conditional indepence on each of the input features given y holds.

(d) the decision rule of predicting a $y = 1$ if $P(y = 1|\mathbf{x}) > P(y = 1|\mathbf{x})$ is equivalent to having a decision boundary defined by $\ln \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})} = 0$.

Based on the above.

$$P(y = 1|\mathbf{x}) = \frac{P(\mathbf{x}|y = 1)P(y = 1)}{P(\mathbf{x}|y = 0)P(y = 0) + P(\mathbf{x}|y = 1)P(y = 1)}$$

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \left[\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=0)P(y=0)}\right]^{-1}}$$

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp \ln \left[\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=0)P(y=0)}\right]^{-1}}$$

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp\left(-\ln \left[\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=0)P(y=0)}\right]\right)}$$

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp\left(-\ln \left[\frac{P(y=1)}{P(y=0)}\right] - \ln \left[\frac{P(\mathbf{x}|y=1)}{P(\mathbf{x}|y=0)}\right]\right)}$$

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp\left(-\sum_i \ln \left[\frac{P(\mathbf{x}_i|y=1)}{P(\mathbf{x}_i|y=0)}\right]\right)}$$

For convenience, let us expand the input to the exponential.

$$\sum_i^d \ln \left[\frac{P(\mathbf{x}_i|y = 1)}{P(\mathbf{x}_i|y = 0)}\right] = \sum_i \ln P(\mathbf{x}_i|y = 1) - \ln P(\mathbf{x}_i|y = 0)$$

$$= \sum_i^d \ln \left[{\theta_{i1}}^{\mathbf{x}_i}(1 - \theta_{i1})^{1-\mathbf{x}_i}\right] - \ln \left[{\theta_{i0}}^{\mathbf{x}_i}(1 - \theta_{i0})^{1-\mathbf{x}_i}\right]$$

After simplifying, this leads to

$$= \sum_i^d \mathbf{x}_i \left[\ln \left(\frac{\theta_{i1}(1 - \theta_{i0})}{\theta_{i0}(1 - \theta_{i1})}\right)\right] + \sum_i^d \ln \left(\frac{(1 - \theta_{i1})}{(1 - \theta_{i0})}\right)$$

which directly corresponds to

$$= \sum_i^d \mathbf{x}_i w_i + w_0 = \mathbf{w}^T\mathbf{x} + w_0$$

Therefore, the expression simplifies to

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp\left(-\mathbf{w}^T\mathbf{x} + w_0\right)}$$

which implies that the generative Naive Bayes binary classifier learns a linear decision boundary $\mathbf{w}^T\mathbf{x} + w_0 = 0$, equivalent to the discrimnatory logistic-sigmoid classifier's decision boundary of predicting $P(y = 1|\mathbf{x}) \geq 0.5$ and $P(y = 0|\mathbf{x}) < 0.5$.

□