Name: **Roll Number**:

CS217 **Artificial Intelligence and Machine Learning** **Quiz-2**

27Mar25 **10.45 AM to 11.25 AM** **20 marks**

(All questions are multiple-choice. There may be multiple correct answers for each question, so you must select all correct options. If your answer is incorrect, including missing any correct options, you will receive -1; otherwise, each correct response earns 2.5 marks.
Below each question, an ANSWER field is provided. You must specify your chosen options only in this field. Marking options with ✓ (tick) or ✗ (cross) beside them will not be accepted. Specifying chosen options anywhere else in the answer sheet will not be considered for grading.

[Q1] Consider designing a linear binary classifier $f(x)=sign(w^Tx+b)$, where $x \in R^2$, on the following training data:

- **Class 1**: {(3,0),(0,3),(3,3)}
- **Class 2**: {(0,0)}

A hard-margin support vector machine (SVM) is used to solve for w and b. Which of the following options is/are correct?

(A) w=[6, 6] and b=1

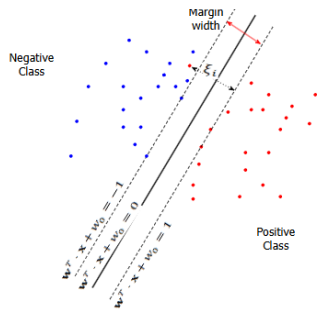
(B) The number of support vectors is 3

(C) The margin is $3/\sqrt{2}$

(D) Training accuracy is 98%

ANSWER:

Solution: (B), (C)

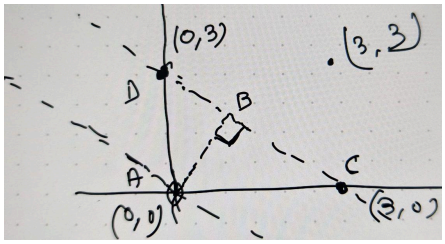


Consider this diagram that was discussed in the class.

(A)w=[6, 6] and b=1 does not satisfy the condition $w^Tx+b <= -1$. Hence, w=[6, 6] and b=1 are not correct parameters.

(B)Support vectors are points that **lie on the margin** (i.e., satisfy the equality in the margin constraints). The support vectors are: (3,0),(0,3),(3,3). Thus, the number of support vectors is 3.

Name: **Roll Number**:



(C)∠ABC = 90 degrees

length(AD) = length(AC)

⇒ ∠ADC = ∠ACD = 45 degrees ⇒ ∠BAC = 45 degrees

⇒ length(AB) == length(BC) (Let say d)

$d^2 + d^2 = \text{length(AC)}^2 \Rightarrow 2d^2 = 9 \Rightarrow$ d (margin width) = $3/\sqrt{2}$

(D)As four points are linearly separable, the accuracy will be 100%

[Q2]

Recall, the original training loss used for standard least squares regression:

$$\text{Loss}(w) = \sum_{i=1}^{n} \left(y^{(i)} - wx^{(i)} \right)^2$$

where $x^{(i)}$ is the input and $y^{(i)}$ is the output for the ith data point in your training set of n data points. Suppose $x^{(i)} \in \mathbb{R}$ and $y^{(i)} \in \mathbb{R}$.

Now, consider a friend proposes a new loss function:

$$\text{NewLoss}(w) = \sum_{i=1}^{n} \left| y^{(i)} - wx^{(i)} \right|$$

Suppose the dataset is fixed, and you can minimize both losses exactly during training. Will these two loss functions always yield the same regression model?

(A) Yes, both loss functions yield the same optimal w because they are minimized under the same dataset.

(B) No, the least squares loss gives more weight to larger errors, while the absolute loss treats all errors equally, leading to different optimal w.

(C) Yes, both loss functions yield the same model, but the convergence behavior during optimization may differ.

(D) No, absolute loss leads to the median estimator of residuals, while squared loss leads to the mean estimator, giving different results in general.

ANSWER:

Solution: (D)

Name: **Roll Number**:

Minimizing these two loss functions yields different optimal solutions because they emphasize errors differently: Least Squares Loss minimizes squared residuals and leads to the *mean estimator*. Absolute Loss minimizes absolute residuals and leads to the *median estimator*.

[Q3] Consider a dataset that is **not linearly separable**. Which of the following statements is **true** regarding the feasibility of SVM solutions?

(A) Hard-margin SVM with a linear kernel never has a feasible solution.

(B) Soft-margin SVM with a linear kernel always has a feasible solution.

(C) Hard-margin SVM with a polynomial kernel always has a feasible solution.

(D) Soft-margin SVM with a nonlinear kernel always has a feasible solution.

ANSWER:

Solution: (A), (B), (D)

Hard-margin SVM requires a dataset to be strictly linearly separable. If the dataset is not linearly separable, there is no feasible solution. Thus, option (A) is correct.

Soft-margin SVM allows some misclassifications using slack variables. It always has a feasible solution because it does not require perfect separation. Thus, option (B), (D) are correct.

A polynomial kernel can help separate some datasets, but not always. There exist datasets that even polynomial transformations cannot make separable. So, option (C) is incorrect.

[Q4] Suppose you have the following training set and have to fit a logistic regression classifier $h_\theta(x)=g(\theta_0+\theta_1x_1+\theta_2x_2)$.

x_1	x_2	y
1	0.5	0
1	1.5	0
2	1	1
3	1	0
2	2	1

Which of the following statement(s) are true?

(A) The cost function J(θ) for logistic regression is always convex, so gradient descent will always converge to the global minimum.

(B) Adding higher-order polynomial terms (e.g., x_1^2, x_2^2, and x_1x_2) could help improve the decision boundary for this dataset.

(C) A linear decision boundary will always be able to correctly classify this dataset.

(D) If we fit a support vector machine (SVM) with a linear kernel on this dataset, it will perform identically to logistic regression.

ANSWER:

Name: **Roll Number**:

Solution: (B)

(A)The cost function J(θ) for logistic regression is convex *only when the dataset is linearly separable*. If the dataset is not linearly separable, it may have multiple local minima, and gradient descent might not always converge to the global minimum.

(B)Adding polynomial terms allows for a more flexible decision boundary, which could better separate the positive and negative examples.

(C)The given data is not linearly separable.

(D)An SVM with a linear kernel and logistic regression do not always yield identical results. SVMs maximize the margin, whereas logistic regression models probabilities using the sigmoid function. Their decision boundaries can differ, especially if the dataset is not perfectly linearly separable.

[Q5] Which of the following changes would commonly cause an SVM's margin to shrink? (*C* is the regularization parameter)

(A) Soft margin SVM: increasing the value of *C*

(B) Hard margin SVM: carefully adding a sample point that violates the margin

(C) Soft margin SVM: decreasing the value of *C*

(D) Hard margin SVM: adding a new feature to each sample point

ANSWER:

Solution: (A), (B)

(A)When C is increased, the penalty for misclassification becomes higher, forcing the algorithm to focus more on minimizing errors than maximizing the margin. This directly results in a narrower margin as the hyperplane adjusts to ensure fewer misclassifications. A large C value makes the SVM behave more like a hard margin classifier, imposing stricter separation requirements

(B)If you add a sample point that violates the margin, a hard margin always shrinks.

In hard margin SVMs, all points must be correctly classified with no violations. When a new point is added that would violate the existing margin, the decision boundary must adjust to maintain strict separation between classes. This adjustment typically requires the margin to shrink to accommodate the new point while maintaining linear separability

(D) If you add a feature, the old solution can still be used (by setting the weight associated with the new feature to zero). Although the new feature might enable a new solution with a wider margin, the optimal solution can't be worse than the old solution.

[Q6] One of the most commonly used kernels in SVM is the Gaussian RBF kernel:

$$k(x_i, x_j) = exp(-||x_i - x_j||^2 / 2\sigma).$$

Suppose we have three points, z_1, z_2 and x. z_1 is geometrically very close to x and z_2 is geometrically far away from x. What is the value of $k(z_1, x)$ and $k(z_2, x)$?

(A) $k(z_1, x)$ will be close to 1 and $k(z_2, x)$ will be close to 0.

(B) $k(z_1, x)$ will be close to 0 and $k(z_2, x)$ will be close to 1.

Name: **Roll Number**:

(C) If σ is very small, then $k(z_1,x)\approx 0$ and $k(z_2,x)\approx 0$, making the kernel ineffective.

(D) The kernel function will return negative values if the bandwidth σ is very small.

ANSWER:

Solution: (A)

(A), (B) If z_1 is **close to** x, then $||z_1-x||^2\approx 0$, so $k(z_1,x)=\exp(0)=1$

If z_2 is **far away from** x then $||z_2-x||^2 \gg 1$, so $k(z_2,x)=\exp(-\text{large value})\approx 0$

(B) is **wrong** because similarity decreases as distance increases, so $k(z_1,x)$ **cannot be close to 0** and $k(z_2,x)$ **cannot be close to 1**.

(C) is **wrong** because a small σ makes the kernel highly sensitive, but it does not necessarily make all values **0**.

(D) is **wrong** because the Gaussian RBF kernel is always **non-negative** (its range is (0,1]).

[Q7] In a Support Vector Machine (SVM) with not linearly separable data, consider the dual formulation with slack variables ξ_i to handle misclassifications. Let x_1,x_2,x_3 be four data points categorized as follows:

1. x_1 is on the correct side of the margin.
2. x_2 is exactly on the margin.
3. x_3 is on the wrong side of the margin but still correctly classified.
4. x_4 is on the wrong side of the decision hyperplane (misclassified).

And the lagrangian is:

$$L(\alpha_1,\cdots,\alpha_N,\mu_1,\cdots,\mu_N) = \min_{\mathbf{w},w_0,\xi}\left(\frac{1}{2}\mathbf{w}^T\cdot\mathbf{w}\right)+C\cdot\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i\cdot\left(-y^i\left(\mathbf{w}^T\cdot\mathbf{x}^i+w_0\right)+1-\xi_i\right) + \sum_{i=1}^{N}\mu_i\cdot(-\xi_i)$$

Given this setup, which of the following statements is **true** regarding their corresponding Lagrange multipliers α_i and slack variables ξ_i?

(A) $\alpha_1=0$

(B) $0<\alpha_2<C$

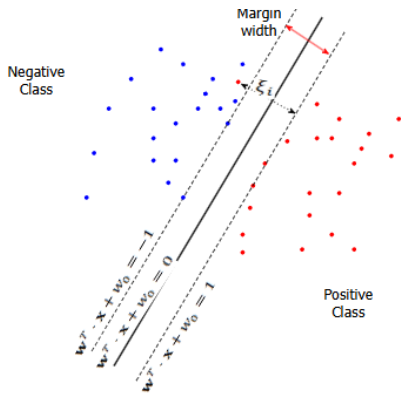
(C) $0<\alpha_3<C$

(D) $\alpha_4=C$

ANSWER:

Solution: (A), (B), (D)

Name: **Roll Number**:



Consider this diagram that was discussed in the class.

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \Rightarrow \boxed{\mu_i + \alpha_i = C}$$ (As we discussed in the class)

(A) It **does not contribute** to the support vectors, so $\alpha_1=0$ (**non-binding constraint**).
(B) If a point is on the margin (case 2), then $\xi_i = 0$, and μ_i can be greater than zero so in general $\alpha_i < C$.
(C) , (D) If a point is on the wrong side of the margin (cases 3 and 4), $\xi_i > 0$ and hence $\mu_i = 0$ (last term of the equation mentioned in the question; μ_i is the Lagrange multiplier for the i^{th} slack variable ξ_i), and hence $\alpha_i = C$.

[Q8] You have a dataset with $n=20$ features and $m=10,000$ examples. After training your logistic regression classifier, you find that it **underfits the training set**, meaning it has high bias. The model performs poorly on both training and cross-validation sets.

Which of the following might be promising steps to improve performance?

(A) Use an SVM with a polynomial kernel.

(B) Increase the learning rate in gradient descent.

(C) Reduce the regularization parameter λ

(D) Reduce the number of training examples to focus on a smaller dataset.

ANSWER:

Solution: (A), (C)

(A) Since underfitting is due to high bias, using an SVM with a **polynomial kernel** introduces non-linearity and allows for a more flexible decision boundary, reducing bias.
(B) Increasing the learning rate affects convergence speed but does **not** resolve underfitting. A model with high bias won't improve significantly just by changing the learning rate.
(C) A high regularization parameter λ makes the model too simple, leading to underfitting. Reducing λ allows the model to fit the data better.

(D) Removing training examples does not help with underfitting. In fact, it could make generalization worse.