دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

یادگیری ماشین

تمرین شماره 3

نام و نام خانوادگی
سید محمد جزایری

شماره دانشجویی
810101399

آذر 1404

فهرست

فهرست شکل‌ها 3
چکیده 4
پرسش 1 - عنوان پرسش 5
1-1. عنوان بخش 5
1-2. عنوان بخش 5
1-3. عنوان بخش 5

4-1. عنوان بخش 5
پرسش 2 - عنوان پرسش..... 6
پرسش 3 - عنوان پرسش..... 6
پرسش 4 - عنوان پرسش..... 7
4-1. عنوان بخش 7
4-2. عنوان بخش 7
4-3. عنوان بخش 7
پرسش 5 - عنوان پرسش..... 7
5-1. عنوان بخش 7
5-2. عنوان بخش 7
5-3. عنوان بخش 9
5-4. عنوان بخش 10
پرسش 6 - عنوان پرسش..... 10
6-1. عنوان بخش 10
6-2. عنوان بخش 10
6-3. عنوان بخش 10
پرسش 7 - عنوان پرسش..... 10
7-1. عنوان بخش 10
7-2. عنوان بخش 11
7-3. عنوان بخش 11
پرسش 8 - عنوان پرسش..... 11
8-1. عنوان بخش 11
8-2. عنوان بخش 12
8-3. عنوان بخش 14
8-4. عنوان بخش 15
8-5. عنوان بخش 15
8-6. عنوان بخش 16
پرسش 9 - عنوان پرسش..... 16
9-2. عنوان بخش 16
9-3. عنوان بخش 16
9-4. عنوان بخش 16
پرسش 10 - عنوان پرسش..... 17
10-1. عنوان بخش 17
10-2. عنوان بخش 17
10-3. عنوان بخش 17
10-4. عنوان بخش 18
10-5. عنوان بخش 18

فهرست شکل‌ها

Fig 1..... 9
Fig 2..... 11
Fig 3. Performance of SVM with and without scaling..... 13
Fig 4. Confusion matrixes of different kernels..... 14
Fig 5. Result of PCA..... 15
Fig 6. Decision boundary..... 17
Fig 7. Decision tree accuracy with different depths..... 18

چکیده

In this assignment we use two famous and powerful classifiers i.e. SVM and decision tree to solve some real life problems and also explore their weaknesses and their reasons and how we could circumvent these faults or at least decrease their effect.

پرسش 1 - عنوان پرسش

1-1. عنوان بخش

1. The dual form uses the dot product of two datapoints in its optimization and this part can be replaced by the kernel function $K(x_i, x_j)$. Hence we needn't compute the dot product of the two datapoints which could even be in infinite-dimensions!
2. As we said in the previous point the dimensionality of the datapoints in the new space could be infinite, and the primal form's time complexity is linear in terms of the dimensions (d), whereas the complexity of the dual form is linear in terms of the number of samples. With all that being said if the dimensionality of the new space is high the dual form is faster.

2-1. عنوان بخش

Because the other datapoints don't really matter in choosing the margins since the margin stops growing and spreading once it hits the first datapoint i.e. the support vector. Furthermore, according to the KKT complementarity conditions, $\alpha_i[y_i(\omega^T x_i + b) - 1] = 0$. For non-support vectors, the constraint is not active (i.e., $y_i(\omega^T x_i + b) > 1$), so to satisfy the condition, the dual coefficient α_i must be zero.

3-1. عنوان بخش

The kernel trick works because the dual SVM optimization problem and the final decision function $f(x) = sign(\sum \alpha_i y_i K(x_i, x) + b)$ only require the evaluation of inner products $\phi(x)^T \phi(z)$. By Mercer's theorem, if a function K(x, z) satisfies certain conditions (positive semi-definite), there exists a mapping ϕ such that $K(x, z) =< \phi(x), \phi(z) >$. This allows us to compute the scalar result of the inner product in the high-dimensional space directly using K, bypassing the need to compute or store the potentially infinite-dimensional vector $\phi(x)$.

4-1. عنوان بخش

Linear: Creates a straight hyperplane. Low capacity, good for linearly separable data, high bias, low variance.

Polynomial: Creates curved decision boundaries. Higher capacity than linear. Higher degree polynomials increase model complexity and the tendency to overfit.

RBF (Gaussian): Creates complex, closed decision regions around data points. Theoretically infinite capacity. A small σ (narrow peak) can lead to severe overfitting (memorizing data), while a large σ acts more like a linear classifier.

پرسش 2 - عنوان پرسش

The RBF kernel $K(x,z) = exp\{-\frac{||x-z||^2}{2\sigma^2}\}$ can be interpreted as an inner product in an infinite-dimensional space because of its Taylor series expansion. Expanding the term $||x-z||^2 = ||x||^2 + ||z||^2 - 2x^Tz \Rightarrow K(x,z) = e^{-\frac{||x||^2}{2\sigma^2}} e^{-\frac{||z||^2}{2\sigma^2}} e^{\frac{x^Tz}{\sigma^2}}$. The last term can be expanded using the Maclaurin series for $e^u = \sum_{i=0}^{\infty} \frac{u^n}{n!}$:

$e^{\frac{x^Tz}{\sigma^2}} = \sum_{i=0}^{\infty} \frac{(x^Tz)^n}{\sigma^{2n} n!}$ Since x^Tz contains products of all combinations of input features, the kernel represents a sum of dot products in a space spanned by all possible monomial features of all degrees, which is an infinite-dimensional Hilbert space.

پرسش 3 - عنوان پرسش

Intuitively these points lie equidistantly from the vertical line $x_1 = 0$ so we expect the answer to be this line. Now the solution. First we form the Lagrangian:

$L_D(\alpha) = \sum_{i=1}^{4} \alpha_i - \frac{1}{2}\sum_{i=1}^{4}\sum_{j=1}^{4} \alpha_i\alpha_j y_i y_j (x_i^T x_j)$. now we have to maximize this term with respect to α subject to: $\sum_{i=1}^{4} \alpha_i y_i = 0 \; and \; \alpha_i \geq 0$. after calculating the dot products we plug in the values and we get:

$L_D(\alpha) = \sum_{i=1}^{4} \alpha_i - \frac{1}{2}(\sum_{i=1}^{4}[2\alpha_i^2] + 2(2\alpha_1\alpha_4 + 2\alpha_2\alpha_3)) = (\alpha_1 + \alpha_4) + (\alpha_2 + \alpha_3)$

$- (\alpha_1 + \alpha_4)^2 - (\alpha_2 + \alpha_3)^2$. Let $S_1 = (\alpha_1 + \alpha_4)$ and $S_2 = (\alpha_2 + \alpha_3)$ so now we have to maximize two independent quadratic equations and the optimum point is $S_1 = S_2 = 0.5$. So overall:

$\alpha_1 + \alpha_4 = 0.5 \; and \; \alpha_2 + \alpha_3 = 0.5$ also from our constraint we have: $\alpha_1 - \alpha_4 + \alpha_2 - \alpha_3 = 0$ and this system of equations has infinite answers but give the symmetry we mentioned earlier they all should be equal hence $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.25$

$\omega = \sum_{i=1}^{4} \alpha_i y_i x_i = [1 \quad 0]^T$ and for b we use the first datapoint:

$b = y_1(\omega^T x_1 + b) = 1 \Rightarrow b = 0.$

Finally: $\alpha = [0.25 \quad 0.25 \quad 0.25 \quad 0.25]$, $\omega = [1 \quad 0]^T$, $b = 0$

And our line is $x_1 = 0$ which is the x_2 axis.

پرسش 4 - عنوان پرسش

1-4. عنوان بخش

The Multinomial Theorem states that for any positive integer n and terms t_1,..., t_m:

$$(t_1 + \ldots + t_m)^n = \sum_{k_1+\ldots+k_m=n} \frac{n!}{k_1!\ldots k_m!} \prod_{j=1}^{m} t_j^{k_j}$$

In our case n is 3 and we have three terms: $1, x_1z_1, x_2z_2$

$$K(x,z) = \sum_{i+j+k=3} \frac{3!}{i!j!k!} (1)^i (x_1z_1)^j (x_2z_2)^k = 1 + 3x_1z_1 + 3x_2z_2 + 3x_1^2z_1^2 + 3x_2^2z_2^2 + 6x_1x_2z_1z_2$$

$$+ x_1^3z_1^3 + x_2^3z_2^3 + 3x_1^2x_2z_1^2z_2 + 3x_1x_2^2z_1z_2^2$$

There are 10 distinct terms in the expansion and therefore the dimensionality of the feature space is 10.

2-4. عنوان بخش

$$\phi(x) = [1, \sqrt{3}x_1, \sqrt{3}x_2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{6}x_1x_2, x_1^3, x_2^3, \sqrt{3}x_1^2x_2, \sqrt{3}x_1x_2^2]^T$$

$$\phi(z) = [1, \sqrt{3}z_1, \sqrt{3}z_2, \sqrt{3}z_1^2, \sqrt{3}z_2^2, \sqrt{6}z_1z_2, z_1^3, z_2^3, \sqrt{3}z_1^2z_2, \sqrt{3}z_1z_2^2]^T$$

3-4. عنوان بخش

$$\phi(x)^T\phi(z) = 1 + 3x_1z_1 + 3x_2z_2 + 3x_1^2z_1^2 + 3x_2^2z_2^2 + 6x_1x_2z_1z_2 + x_1^3z_1^3 + x_2^3z_2^3$$

$$+ 3x_1^2x_2z_1^2z_2 + 3x_1x_2^2z_1z_2^2$$

پرسش 5 - عنوان پرسش

1-5. عنوان بخش

$$H(y) = -P(y=0)log_2P(y=0) - P(y=1)log_2P(y=1) = -0.5log_2(0.5) - 0.5log_2(0.5)$$

$$= 1.0(bit)$$

2-5. عنوان بخش

X_1:

$$X_1 = A: P(y=0|X_1=A) = \frac{2}{3} \; and \; P(y=1|X_1=A) = \frac{1}{3}$$

$$H(y|X_1=A) = -(\frac{1}{3}log(\frac{1}{3}) + \frac{2}{3}log(\frac{2}{3})) \approx 0.918$$

$X_1 = B: P(y = 0|A) = \frac{1}{3}\ and\ P(y = 1|A) = \frac{2}{3}$

$H(y|X_1 = B) = -(\frac{2}{3}log(\frac{2}{3}) + \frac{1}{3}log(\frac{1}{3})) \approx 0.918$

Weighted entropy: $0.5(0.918) + 0.5(0.918) = 0.918$

$IG(y; X_1) = 1 - 0.918 = 0.082$

X_2:

$X_2 = C: P(y = 0|X_2 = C) = 1\ and\ P(y = 1|X_2 = C) = 0$

$H(y|X_2 = C) = -(1log(1) + 0log(0)) = 0$

$X_2 = D: P(y = 0|D) = 0\ and\ P(y = 1|D) = 1$

$H(y|X_2 = D) = -(0log(0) + 1log(1)) = 0$

Weighted entropy: $0.5(0) + 0.5(0) = 0$

$IG(y; X_1) = 1 - 0 = 1$

The algorithm will choose X_2 since it has greater information gain.

3-5. عنوان بخش

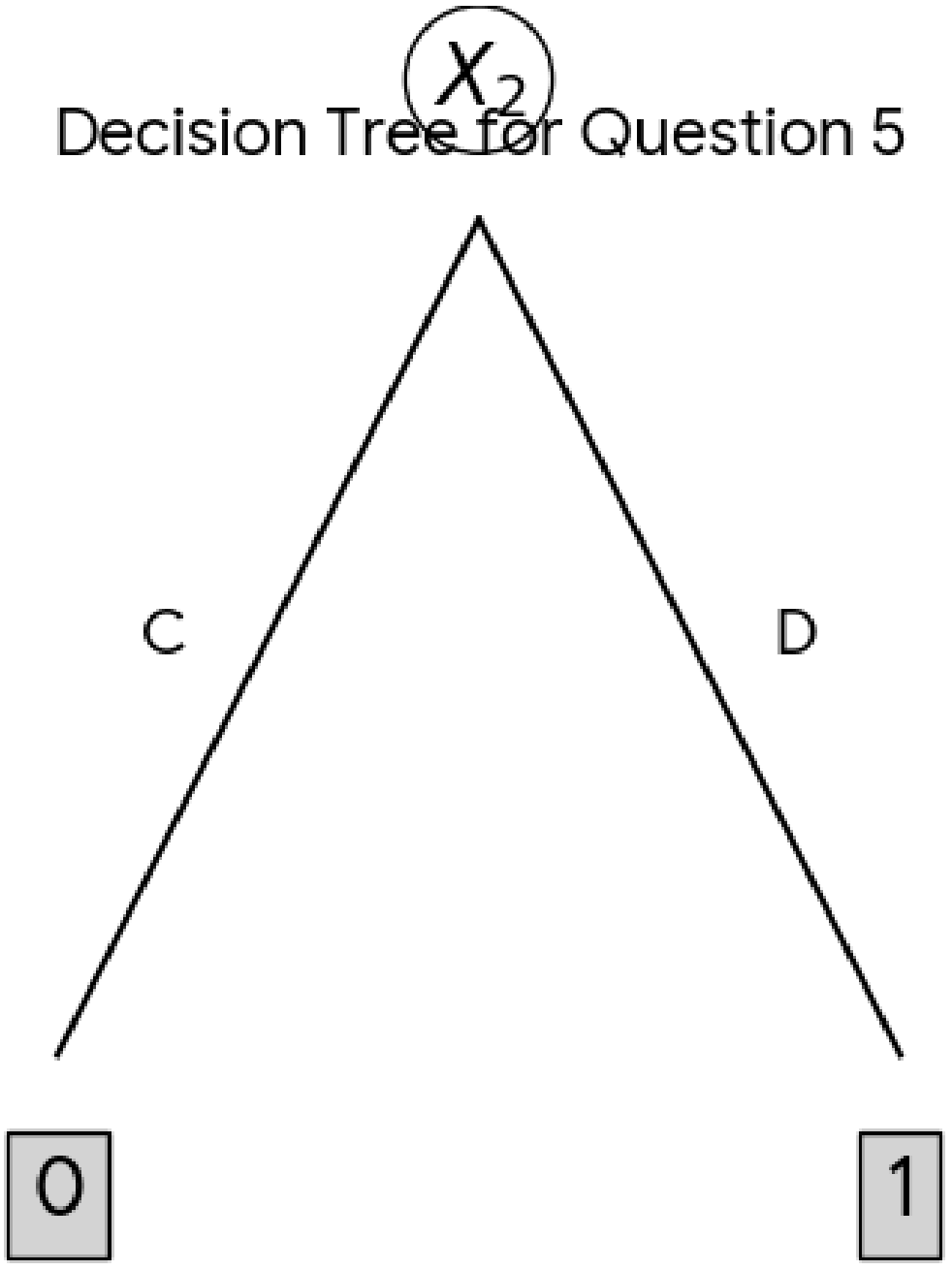

Fig 1.

Yes it does reach 100% accuracy on train data.

4-5. عنوان بخش

Top-down is greedy. A bottom-up approach (like pruning back a full tree or clustering) might produce the same tree here because the data is perfectly separable by a single feature. However, generally, they are not necessarily identical. Bottom-up methods consider global structure or complexity constraints that greedy top-down splits miss.

پرسش 6 - عنوان پرسش

1-6. عنوان بخش

$H'(p) = ln(1 - p) - lnp$

$H''(p) = -\frac{1}{p(1-p)} < 0 \; for \; 0 < p < 1$ Thus H(p) is strictly concave.

2-6. عنوان بخش

Information Gain is $H(S) - \sum \frac{|s_v|}{|s|} H(s_v)$.

This is equivalent to $H(E[Y]) - H(E[Y|X])$.

By Jensen's Inequality for a concave function f (Entropy): $f(E[x] \geq E[f(x)]$.

Therefore, $H(Average\ Posterior) \geq AverageH(Posterior)$.

$H(S) \geq \sum w_i H(s_i) \Rightarrow IG \geq 0$.

3-6. عنوان بخش

ID3 is a greedy hill-climbing algorithm. It chooses the locally best split. Scenario: The XOR problem (Question 7). A feature might have 0 gain individually (looks like noise), but in combination with another feature, it provides perfect classification. ID3 sees 0 gain at the root and might discard the feature or stop, failing to find the global optimum which requires looking ahead 2 steps.

پرسش 7 - عنوان پرسش

1-7. عنوان بخش

$P(y = 0) = P(y = 1) = 0.5$

$H(y) = 1$

Split based on X_1:

$H(y|X_1 = 0) = H(y|X_1 = 1) = 1(bit)$

$IG(y; X_1) = 0$

And the same goes for X_2. Since Gain is 0 for all features, ID3 sees no value in splitting. It might stop (returning majority class node with 50% error) or pick randomly.

2-7. عنوان بخش

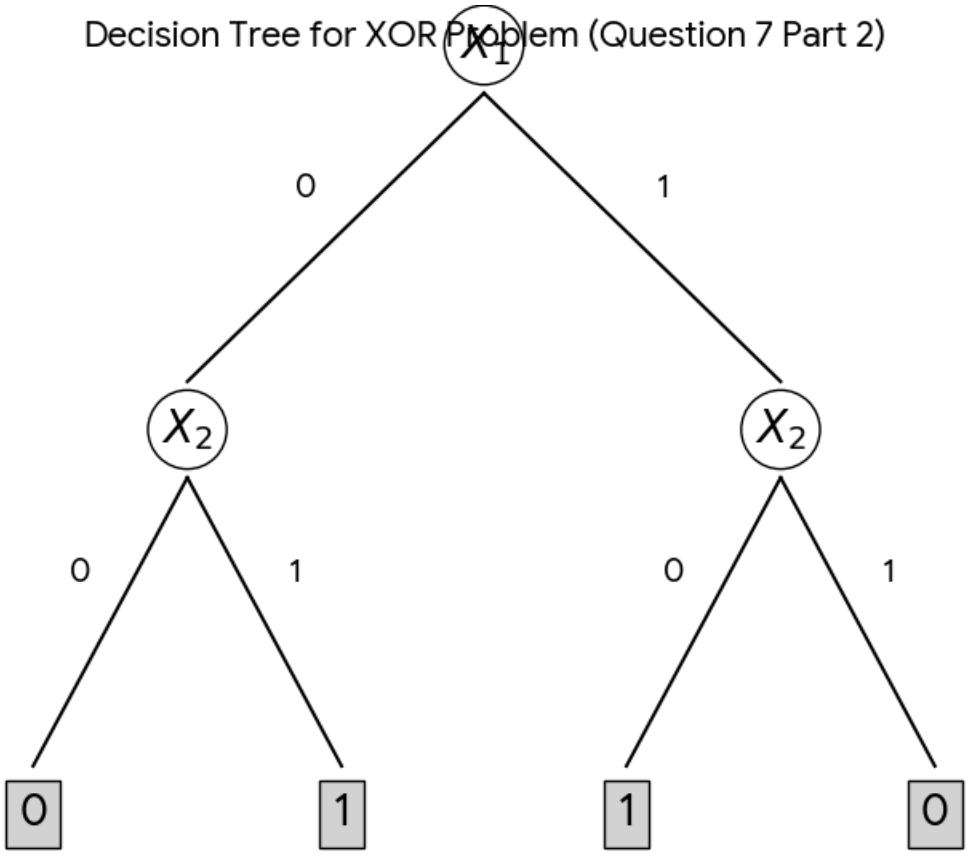

Fig 2.

3-7. عنوان بخش

(a) Preference Bias: The algorithm prefers shorter trees and places high information gain attributes close to the root. It does not strictly rule out complex trees (Restriction Bias), but simply searches the space in an order that prefers simple ones.

(b) Failure Explanation: The failure on XOR shows that ID3 doesn't check all possible trees (hypothesis space). It uses a heuristic (Info Gain). Because it prefers high gain *immediately* at the root, it fails to see that splitting on X_1 (Gain 0) is necessary to unlock the perfect split on X_2 later. BFS-ID3 (Breadth-First Search) would find it by exploring all depth-2 trees, but standard ID3 is greedy depth-first.

پرسش 8 - عنوان پرسش

1-8. عنوان بخش

SVM attempts to maximize the margin distance between classes. If one feature has a range of 0–1000 (e.g., Insulin) and another 0–1 (e.g., Pedigree Function), the distance calculation ($||x - z||^2$) will be completely dominated by the larger feature. The SVM would effectively ignore the smaller feature. Scaling ensures all features contribute equally to the decision boundary.

2-8. عنوان بخش

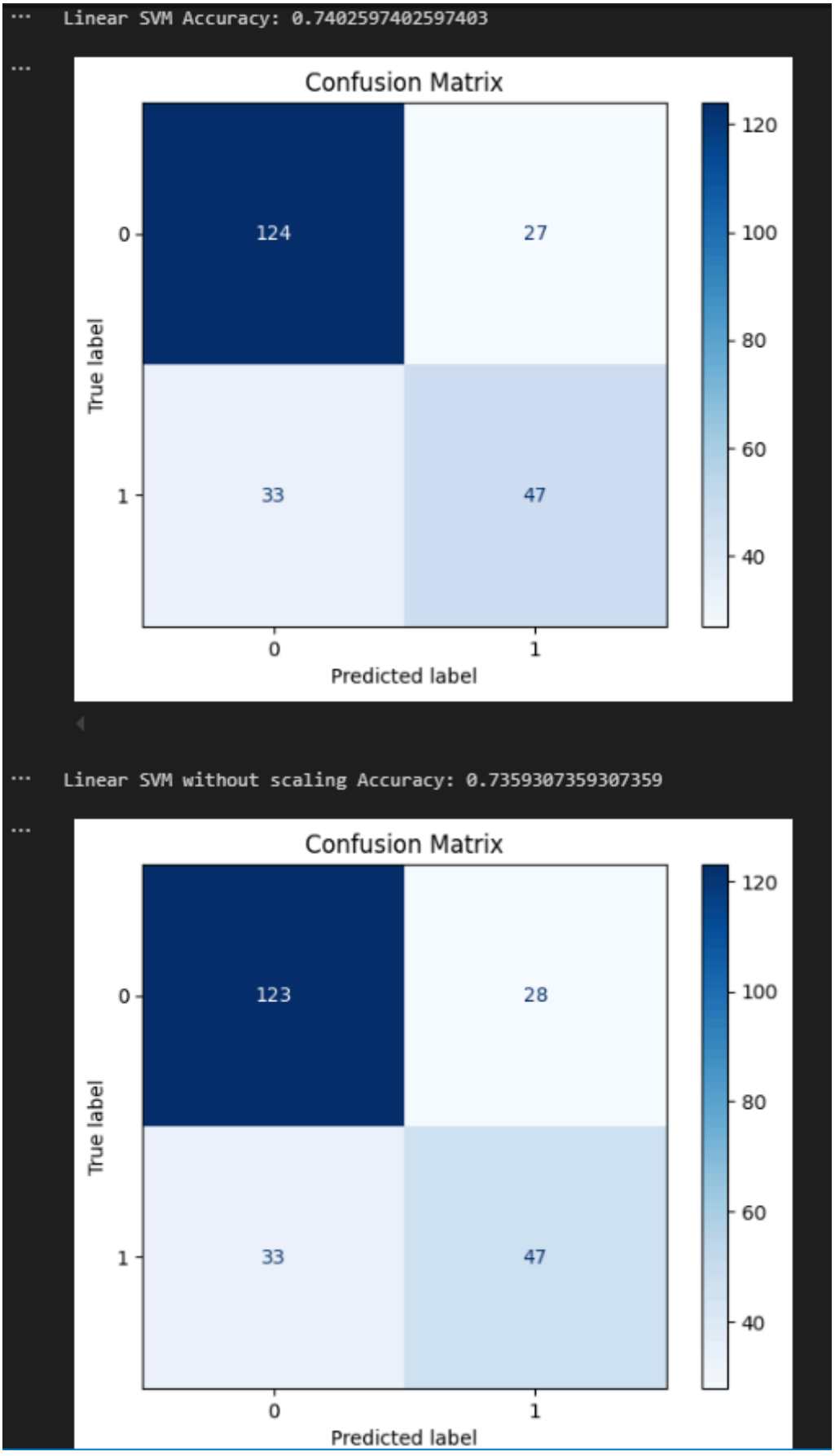

Fig 3. Performance of SVM with and without scaling

In this specific instance the numerical accuracy change (0.44%) is not strictly significant in terms of raw predictive power.

But theoretically and mechanically normalization is fundamentally important for SVMs for the following reasons:

- Distance-Based Margin Calculation: SVM maximizes the margin between the decision boundary and the support vectors using Euclidean distance. If one feature has a range of 0-1000 (e.g., Insulin) and another has a range of 0-1 (e.g., DiabetesPedigreeFunction), the distance metric will be dominated by the larger values. This causes the SVM to effectively ignore the influence of the smaller-scaled feature.
- Numerical Stability and Convergence: Support Vector Machines solve a quadratic programming problem. When features have wildly different scales, the optimization surface becomes very "steep" in some directions and flat in others, which can lead to longer training times or numerical instability during solver convergence.
- Support Vector Selection: Feature scaling ensures that the support vectors are determined by the intrinsic relationship of the features to the target label rather than the arbitrary units (magnitude) in which the features are measured.

3-8. عنوان بخش

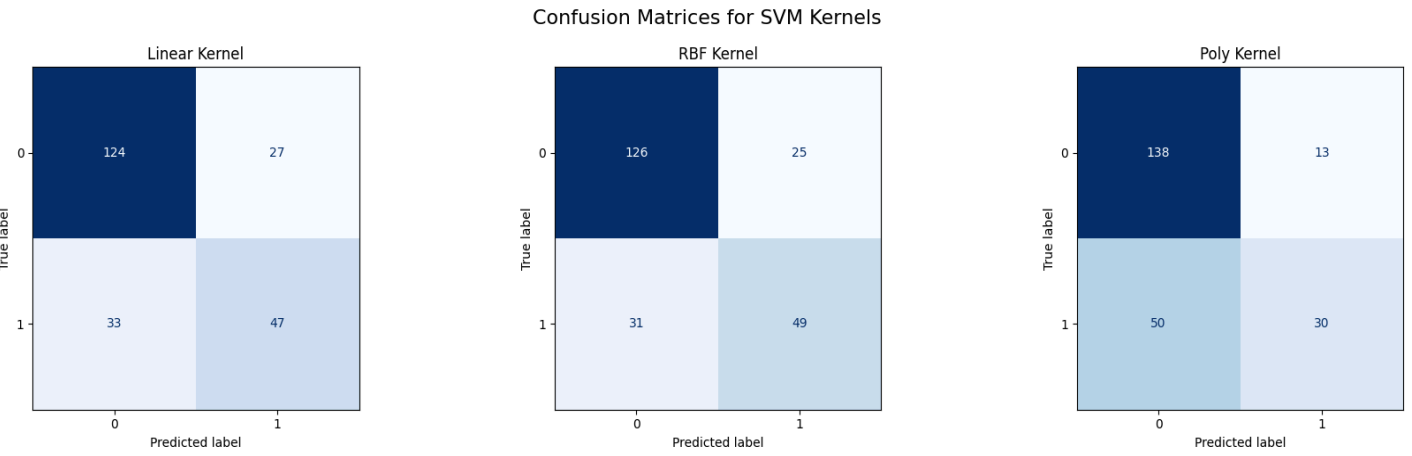

Fig 4. Confusion matrixes of different kernels

The RBF Kernel is the best performing model for this dataset for the following reasons:

1. Highest Recall (Sensitivity): In medical diagnosis (detecting diabetes), Recall is the most critical metric. You want to minimize False Negatives (missing a sick patient).
    - The RBF kernel missed the fewest cases (**31**), compared to Linear (**33**) and Poly (**50**).
    - It correctly identified **61.3%** of the positive cases, which is the highest among the three.

2. Highest Overall Accuracy: It achieved an accuracy of **75.8%**, correctly classifying more patients overall (175 total correct) compared to Linear (171) and Poly (168).
3. Non-Linearity: The slight improvement over the Linear kernel suggests that the decision boundary between diabetic and non-diabetic patients is non-linear. The RBF kernel successfully captured some of these more complex patterns.

4-8. عنوان بخش

Best Params: {'C': 10, 'gamma': 0.001}, Best Score: 0.77466251298027

This is the output of the best parameters.

A very large C strictly penalizes any misclassification. This forces the SVM to create a complex, wiggly boundary to fit every single training point correctly (Hard Margin). This reduces training error to near zero but drastically increases variance, leading to overfitting and poor generalization on new data.

5-8. عنوان بخش

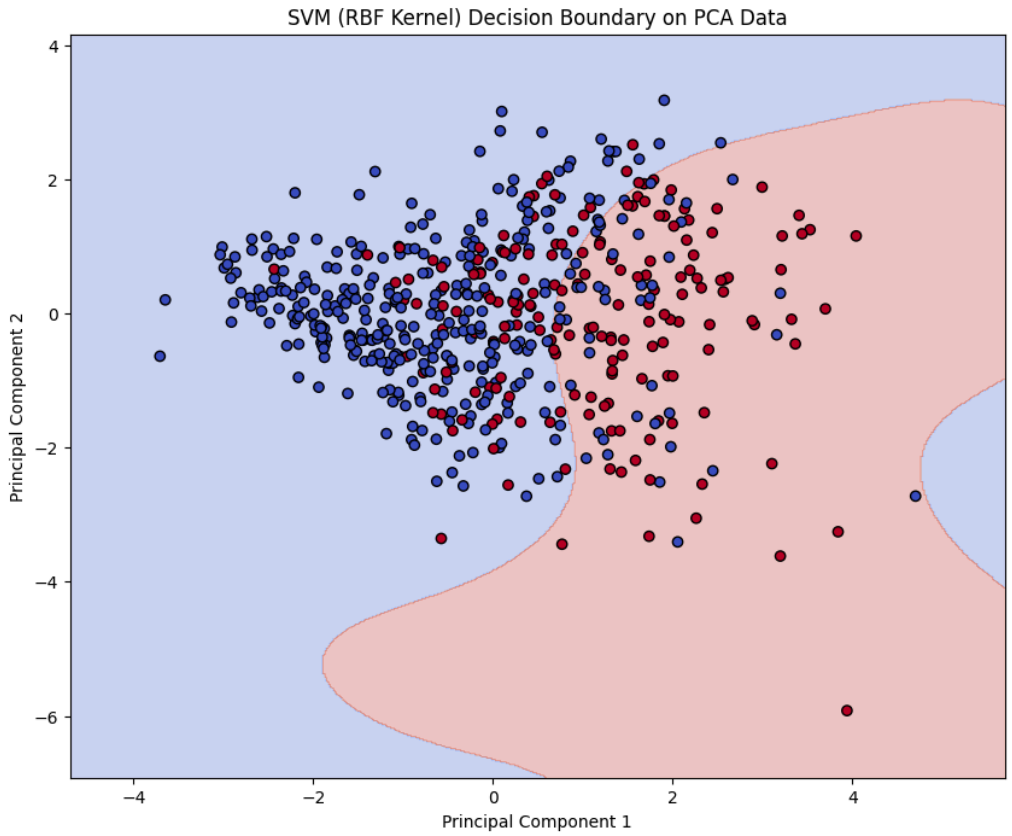

Fig 5. Result of PCA

Our accuracy now is 71%. PCA typically lowers the accuracy compared to the full 8-dimensional model.

Why? PCA reduces dimensions by discarding components with lower variance. While the first two components capture the most variance, they do not capture all of it. The information

lost in the discarded 6 dimensions often contains subtle patterns necessary for distinguishing between the classes, leading to a loss in predictive power. However, the trade-off is that we can now visualize the data and the decision boundary in 2D.

6-8. عنوان بخش

Since 268 have diabetes and 500 don't, the class is imbalanced. Now recall has surged to 0.78 from 0.61 which is a great improvement.

Significant Improvement in Recall: The primary goal of handling class imbalance in a medical dataset is to stop the model from ignoring the minority class (sick patients). This means our model is now correctly identifying 78% of the diabetic patients, whereas previously it missed nearly 40% of them.

The Trade-off: This improvement comes at a cost. The Precision for Class 1 dropped (from 0.66 to 0.58), meaning there are more false alarms (healthy people predicted as diabetic). Consequently, the Overall Accuracy dropped slightly (from 76% to 73%) because the model is now sacrificing some accuracy on the majority class (Class 0 recall dropped) to ensure it catches the minority cases.

Conclusion: In the context of medical diagnosis, this is a successful result. It is generally considered better to have a slightly lower overall accuracy but a higher sensitivity (recall) so that fewer positive cases go undetected.

پرسش 9 - عنوان پرسش

2-9. عنوان بخش

Optimal w : [1.6032232 0.971575]

Optimal b : -0.7688348265267029

3-9. عنوان بخش

QP-SVM Accuracy: 0.90

4-9. عنوان بخش

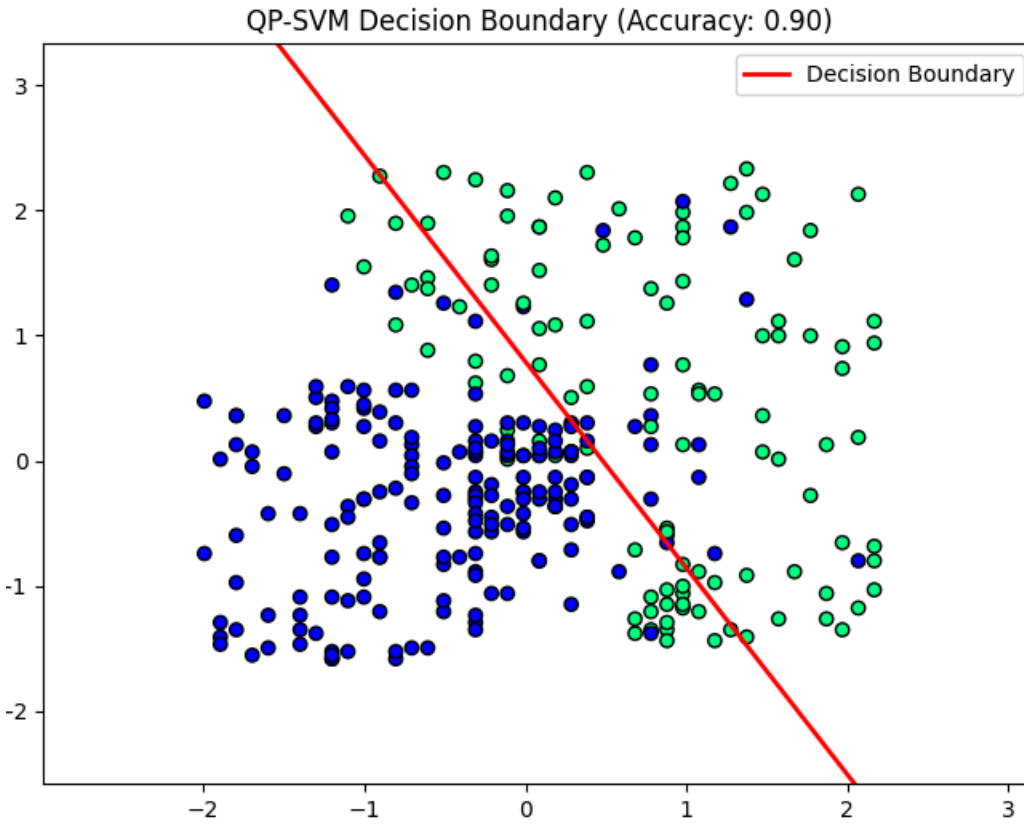

Fig 6. Decision boundary

پرسش 10 - عنوان پرسش

10-1. عنوان بخش

Information Gain for Day_ID: 0.9815

Reason: Day_ID is unique for every row, resulting in pure leaves (Entropy 0) and maximal Gain. It overfits perfectly but generalizes poorly.

10-2. عنوان بخش

--- Information Gain at Root ---

Outlook: 0.3431

Humidity: 0.0539

Wind: 0.0382

Temperature: 0.0239

Feature with highest gain: Outlook

10-3. عنوان بخش

(30.05, 0.02392309702606632)

The first element is the threshold and the second element is the information gain.

4-10. عنوان بخش

Accuracy: 0.94

Precision: 0.95

Recall: 0.95

5-10. عنوان بخش

--- Overfitting Analysis ---

Tree Depth on Clean Data: 11

Tree Depth on Noisy Data: 14

Observation: The tree on noisy data is likely deeper because it tries to memorize the contradictions.

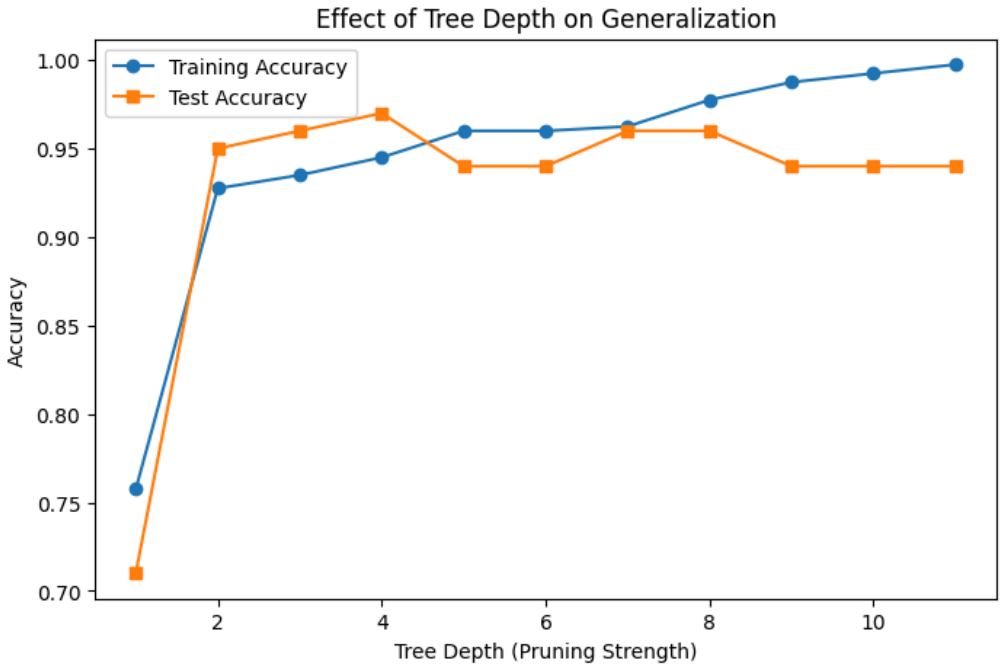

Fig 7. Decision tree accuracy with different depths

We can see very clearly in this plot that with increasing the depth of the tree our training accuracy gets better and better. But that's not the case for the test accuracy; it actually starts decreasing drastically after depth 8 because it can't generalize. It just memorizes the data which leads to overfitting!