CM146: Quiz 8 Solutions

March 10, 2023

1 PAC Learnability 1

1.1

The PAC learnability general bound concerns hypotheses classes which contain a finite number of hypotheses. Infinite hypotheses classes must be cateogrized using other ways such as the VC Dimension, which is not part of the PAC learnability general bound. False.

1.2

The hypothesis class contains 100 hypotheses ($|H| = 100$). We are interested in an error rate of 0.1 ($\epsilon = 0.1$). The probability of the bound should be set to 0.95 ($\delta = 1 - 0.95$, $\delta = 0.05$). Plugging the values into the general bound, we get:

$$m \geq \frac{1}{0.1}(ln(100) + ln(\frac{1}{0.05})) \tag{1}$$

2 VC Dimension 1

It is important to note at least in the context of this problem that ”any” does not mean the same thing as ”all.” To satisfy any (ie. any one), some or at least one condition needs to be satisfied.

1. VC >= D if the classifier can shatter any labelling of D points. True. If there is one example that can be shattered than D is the minimum possible VC.
2. VC = D if the classifier can shatter all labelling of D points. False. This only proves that D is the minimum possible VC, not equals.
3. VC < D if the classifier can shatter any labelling of D points. False. The ability to shatter shows the lower bound, not the upper bound.
4. VC = D - 1 if the classifier cannot shatter any labelling of D points. False. There is no guarantee that D - 1 will be the VC if no D points can be shattered (D - 1 may also be impossible to shatter).

3 VC Dimension 2

True. Increasing the degree of the polynomials increases their complexity (consider $f(x) = (x+1)(x-1)$ vs $f(x) = (x+1)(x-1)(x+2)$, the cubic function crosses the x-axis more times than the squared function). This implies that the VC will increase when higher degree polynomials are included in the hypothesis class. One further clarification is that the degree discussed in this problem referred to the degree of the polynomial functions, not the dimensionality of the data.

4 Kernels 1

The statement that ”Any function that is bivariate $f(u, v)$ is a valid kernel function.” is False since bivariate functions are not guaranteed to be Positive Semi Definite.

5 Kernels 2

False. The addition rule for kernels presumes that the kernels added are indeed kernels. While scalars may be treated as kernels, -1 is not a valid kernel (the ϕ function cannot be defined; it is not PSD). Therefore there is no guarantee that the suggested operation will yield a valid kernel.

6 SVM 1

$\xi_1 \geq 1$ was the correct answer. A misclassification for x_1 suggests that $y_1(w^T x_1 + b) \leq 0$ because the signs of the label and the classification will not match. Then given the constraint

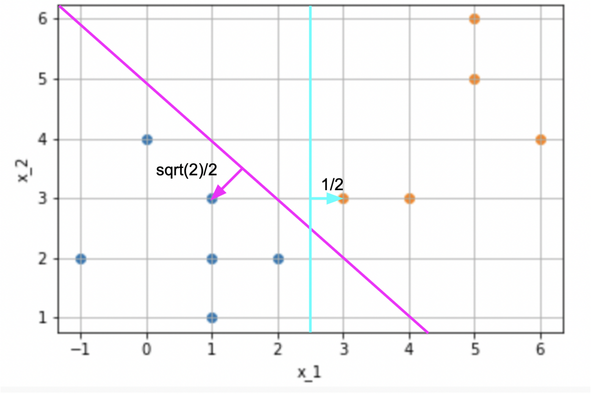
$$y_1(w^T x_1 + b) \geq 1 - \xi_1 \tag{2}$$

only by increasing ξ_1 to at least 1 or greater will allow the misclassification.

7 SVM 2

$5 - x_1 - x_2 = 0$ is the boundary chosen by the SVM. The question asks about the ”boundary” which is the decision boundary of the SVM. Visually, the boundary should show positive points on one side and negative points on the other. This is related to but not the same as the formula for classifier itself which defines a vector normal to the decision boundary (as indicated using the variable w or θ). In other words, the equation for the boundary and the classifier are not the same.

The definition of the margin must be considered to distinguish between $x_1 - 2.5 = 0$ and $5 - x_1 - x_2 = 0$ which both pass through the blue and orange points. $x_1 - 2.5 = 0$ may be appealing because it is further away from some blue points, but the margin is the distance to the closest point (margin of 0.5). $5 - x_1 - x_2 = 0$ has a larger margin by going through the closest points diagonally ($\sqrt{2}/2$). Therefore, the SVM would result in the boundary of $5 - x_1 - x_2 = 0$.



8 SVM 3

1. The α variables must be obtained for the testing examples to predict using the Kernel SVM. False. The alpha variables only exist for the training examples. There are no alphas defined over testing examples.

2. We can predict the label of a new sample using the kernel function and the training data. Partially True. The alpha variables, the labels, and the bias are also needed to make predictions. Points were given based on justification.

3. If we apply kernel functions, non-separable data may become separable. True. The purpose of kernel functions were to perform transformations on the data.

4. A valid kernel function should have a positive-semidefinite kernel matrix. True. By definition.

9 Metric Choice

High recall, but low precision is the correct answer. Since Recall appears for all four choices (Sensitivity = Recall), it is important to check if Recall would be desirable in this setting. Given that Recall = $\frac{TP}{TP+FN}$, a high Recall is obtained when there are low False Negatives and high True Positives, which can reduce missed diagnoses and increase the detection of the disease. Low Recall would imply higher amounts of the opposite, which will lead to many missed diagnoses.

Given that Precision = $\frac{TP}{TP+FP}$, as long as the True Positives is not lower (which may be implied by high Recall), lower Precision can be caused by more False Positives which is not a critical concern.

Given that Specificity = $\frac{TN}{TN+FP}$, neither True Negatives nor False Positives inform directly the ability for a classifier to detect the disease when a patient actually has it (the TP, FN cases).