CSDS 440: Assignment 3

Shaochen (Henry) ZHONG, sxz517

Mingyang TIE, mxt497

Due on 09/25/2020, submitted early on 09/18/2020

Problem 10

It depends on the task and how is the performance of the model being "worsen" on test data than training data. Say we have a task to detect fire of a building so that less people get hurt, and the model performance is lower on test data than training data due to having a lot of false positives. Such model may still be beneficial as the cost of having a false negative is a lot more expensive than having a false positive in this particular task. And even though the model might be overfitting by definition, it might perform better than a model that is less overfit but has more false negative.

The other possible scenario include but not limited to:

- When we want to test the upper capacity of our model or we plan to compare the fitting capacity of some certain kinds of models, we usually keep models overfit for the data.
- When there is no noise in the data, we want our model to fit the pure data as precisely as possible. Under this circumstance, overfit is beneficial. For example, using least square method to fit the curve of an ideal polynomial.

Problem 11

A preference bias is an inductive bias where some hypothesis are preferred over others.

- Pros: By using prior knowledge, it allows the learner to work within a complete hypothesis space that is assured to contain the unknown target function.
- Cons: We must have correct prior knowledge about hypothesis space. Otherwise, we may miss the target function.

A restriction bias is an inductive bias where the set of hypothesis considered is restricted to a smaller set.

- Pros: By restrict hypothesis space, we may need less time and effort to search entire hypothesis space.
- Cons: It strictly limits the set of potential hypotheses, which may bring the possibility of excluding the unknown target function altogether.

Problem 12

No. This is not a goth methodology due to an effective concept usually takes a reasonably large amount of training to learn. However by having an equal-size training and evaluation sets, the training set is likely not large enough – or at least not as effective having a larger training set.

Also because of the equal-sized division, the examples in the training set of during an iteration can be very different to another iteration. This inconsistency will increase the difficuty for person X to analyse wheather it is the problem on training data or the model itself, should there ever be any undesired/unstable performance measures.

At the same time, some examples might always being divided to the training test during each iterations, so the model will therefore never be able to evaluate its ablity of predicting these examples; and this for sure lower the realiability of the performance measure of the model. In general, a N-fold approach will be preferred.

Problem 13

Because ROC graph is ploting TP Rate $= \frac{TP}{TP+FN}$ against FP Rate $= \frac{FP}{FP+TN}$. As we are lowering the classification threshold, more examples will be classfied as *Positive*. This implies there will be more TP and FP (for a typical model) as the threashold being lower – and since both $TP + FN$ and $FP + TN$ are constant for all time – we will have a larger numerators on the same denominators. Which will result in an increase on both axises and the statament is therefore proven.

Problem 14

Let R denotes the examples that are being classfied as *Positive*, and T denotes the true posive cases. We have:

$$\begin{aligned}
P(R) &= P(R \mid T)P(T) + P(R \mid T^c)P(T^c) \\
&= P(R \mid T)(1 - P(T^c) + P(R \mid T^c)P(T^c) \\
&= P(R \mid T) - P(R \mid T)P(T^c) + P(R \mid T^c)P(T^c) \\
&= P(R \mid T) + [P(R \mid T^c) - P(R \mid T)] \cdot P(T^c) \\
P(R) - P(R \mid T) &= [P(R \mid T^c) - P(R \mid T)] \cdot P(T^c)
\end{aligned}$$

We know that there must be $P(R) = P(R \mid T)$ as random guessing is an independent variable. We also know we should have $P(T^c) > 0$ for being a meaningful task. Subsituting these into the above equation, we have $0 = P(R \mid T^c) - P(R \mid T) \implies P(R \mid T^c) = P(R \mid T)$. This implies that the TP Rate is the same as the FP Rate and therefore $x = y$, and the ROC graph for a random gussing classifier will therefore be a diagonal line.