MTH 9899: Machine Learning
Bo Feng
Spring 2015

Problem Set # 3
Solution

1. Problem 1

   By definition, the unadjusted Sharpe ratio is the return over its observed volatility. Hence

   $$SR \simeq \frac{E(Y^*)}{\text{std}(Y^*)} = \frac{\lambda\sigma_Y}{\sqrt{1+n^2}\sigma_Y} = \frac{\lambda}{\sqrt{1+n^2}} \simeq \lambda R,$$

   where in the last equality we used the fact that $R^2 \simeq maxR^2 = 1 - \frac{n^2}{1+n^2}$.

2. Problem 2

   a) The diagonal line $y = x$ represents the strategy of randomly guessing a class, i.e., the Random Classifier Line. This strategy can be understood as guessing the positives with probability $p \in [0, 1]$ irrespective of any other information. Consequently, given that the outcomes are all positive, we have probability p of guessing it correctly. The true positive rate is p. Similarly, given all negatives, we still have probability p of guessing it positive, which leads to a false positive rate p. Thus a random classifier will produce a ROC point that "slides" back and forth on the diagonal based on the frequency with which it guesses the positive class.

   b) The Euclidean distance from the given classifier to the random classifier line is the distance from the point (fp, tp) to (fp, fp), i.e., $tp - fp$. The Youden Index $J = sensitivity + specificity - 1$, where sensitivity is the true positive rate tp and specificity is the true negative rate $tn = 1 - fp$. Hence $J = tp + (1 - fp) - 1 = tp - fp$, equivalent to the Euclidean distance from the given classifier to the random classifier line.

   c) Since $DistAcc = 1 - \sqrt{\frac{1}{2}[(1-tp)^2 + fp^2]}$ and $J = tp - fp$, the two types of measure are not necessarily equal. An example would be setting $(fp, tp) = (0.3, 0.5)$.

   d) Consider two classifiers A and B such that $(fp, tp)_B = (0.2, 0.3)$ and $(fp, tp)_A = (0.3, 0.6)$. By simple calculation, $DistAcc(B) = 0.4852$ and $DistAcc(A) = 0.6464$. Hence $DistAcc$ ranks A as better than B. Note that $CostAdjEuclDistAcc = 1 - \sqrt{W(1-tp)^2 + (1-W)fp^2}$. Setting $W = 0.1$ will penalize more on fp. In this case, the $CostAjdEuclDistAcc$ for A and B are 0.6886 and 0.7085 respectively, reverting the rank of A, B. This happens because the cost is heavily incurred on false positive rate fp. With a larger fp for A, it is unfavorably ranked.

3. Problem 3

   When $L > P$, it simply implies that $\theta = \frac{L}{L+P} = \frac{1}{1+P/L} > 0.5$. Hence the classifier threshold should be set higher than 50%.

4. Problem 4

   1) Expected profit comes from the profitable trades on which we take action. In other words, if a trade is profitable and we classify it as positive, then we make a profit. The number of correctly

classified positive predictions is TP. Hence the expected profit is $TP * P$. In contrast, expected loss not only comes from the incorrectly classified negatives (FP), each of which incurs a loss of L, but also from the missed opportunities ($FN * P$).

2) The relation $\frac{TP-FN}{FP} > \frac{L}{P}$ makes sense in view of the following aspects. As TP increases, we are able to make more correct profitable decisions, the inequality will be more easily satisfied. The same is true as FN decreases, which means that the number of opportunities we miss is very small. Also, a small FP implies small loss, making the inequality easy to hold. As the loss amount L increases, $\frac{TP-FN}{FP}$ has to be very large. Either a large TP or a small FN or TP is required.

3) This relation is imposed on the actual numbers of true positives, false negatives and false positives, as opposed to the corresponding rates used in Youden Index. It also involves the actual losses and profits. Therefore such measure is more practical and intuitive. Youden Index is more tailored to cost adjustment and model comparison.

4) True negatives mean that the trades are not profitable and we also correctly classify them as negative. This does not incur any loss when making such a decision. Hence TN does not enter the expression for classifier performance.

5) When $L > P$, false positives are more costly than false negatives. The expected loss is a linear combination of FP and FN, with weights L and P respectively. $L > P$ means we put more weight/cost on FP.

5. Problem 5

   The diagonality measure is not always useful as the sole criterion for comparison.

   In this case, the sum of diagonal entries for both classifiers are equal to 5288. Thus we cannot tell which classifier is better.

   The number of pairwise binary classifiers we should analyze is

$$\binom{5}{2} = 10.$$

   We cannot treat all such pair-wise comparisons the same way. Some misclassification (for example, +2 predicted as -2) are more serious and might incur more cost than some other misclassicication (for example, +2 predicted as +1).

   If we have N classes instead of 5, $\binom{N}{2}$ pair-wise comparisons should be made.

   The analysis method should be similar if asymmetric transaction cost is added, but actual negative profit predicted as positive (-1, -2 predicted as +1, +2) is a more severe mistake than the other way around. This kind of misclassification should be given more weight.

6. Problem 6

   (a) Non-null Attribute Mean (or Median) Imputation:

   advantage: sample mean for the variable is not changed; easy to implement. It is easily parallelizable.

   disadvantage: mean imputation attenuates any correlations involving the variable(s) that are imputed. This method is not very accurate.

   (b) Hot-Deck Imputation:

   advantage: conceptual simplicity, maintenance and proper measurement level of variables. This method gives fairly accurate, relatively cheap to implement, and parallelizable.

   disadvantage: difficulty in defining what is "similar".

(c) KNN Imputation:

advantage: KKN Imputation can predict both discrete attributes and continuous attributes. There is no necessity for creating a predictive model for each attribute with missing data. It gives more accurate result than Hot-Deck, and is parallelizable.

disadvantage: whenever the k-nearest neighbour looks for the most similar instances, the algorithm searches through all the data set. This limitation can be very critical for KDD, since this reearch area has, as one of its main objectives, the analysis of large databases.

(d) LSR Imputation:

advantage: LSR method makes most use of the information from the original data sets. It gives more accurate result than Hot-Deck, and is parallelizable.

disadvantage: expensive to implement compared with other methods.

(e) SVD Imputation:

advantage: it makes use of all available data.

disadvantage: expensive to implement, and is suitable only for small to medium-sized data sets. It is not easily parallelizable due to the relative norm computation at each iteration. SVD imputation relies on the ad-hoc parameters Δ and δ. SVD method is not self-starting, relying on another imputation method to bootstrap the first iteration.

7. Problem 7

(a) Support of $X \Rightarrow Y$ is

$$\frac{200,000}{1,000,000} = 20\%.$$

This is higher than the support threshold.

(b) Confidence of $X \Rightarrow Y$ is

$$\frac{200,000}{500,000} = 40\%.$$

This is higher than the confidence threshold.

(c)According to the Support-Confidence Framework, this is an interesting association rule.

However I don't think it is a good idea. $P(X \cap Y) = P(X)P(Y)$ in this case, so X and Y are independent. X and Y are actually not related to each other.

(d) Support of $X \Rightarrow Y$ is

$$\frac{2,000}{1,000,000} = 0.2\%.$$

This does not meet the support threshold.

(e) Confidence of $X \Rightarrow Y$ is

$$\frac{2,000}{500,000} = 0.4\%.$$

This does not meet the confidence threshold.

(f) According to the Support-Confidence Framework, there is no interesting association rule.

However I don't think it is a good idea. X and Y occurs together less frequently than the independent situation. I assume there is negative relationship between X and Y.