Solutions Week-05: Classification Metrics

1. **Basic Confusion Matrix & Accuracy:**

	Predicted Negative	Predicted Positive
Actual Negative	35 (TN)	15 (FP)
Actual Positive	10 (FN)	40 (TP)

   Accuracy $= \frac{TP+TN}{\text{Total}} = \frac{40+35}{40+10+15+35} = \frac{75}{100} = 0.75$

2. **Precision and Recall:** Precision $= \frac{TP}{TP+FP} = \frac{40}{40+15} = \frac{40}{55} \approx 0.727$ Recall $= \frac{TP}{TP+FN} = \frac{40}{40+10} = \frac{40}{50} = 0.80$

3. **F1-Score:** F1 $= 2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} = 2 \times \frac{0.727\times0.80}{0.727+0.80} = 2 \times \frac{0.5816}{1.527} \approx \frac{1.1632}{1.527} \approx 0.762$

4. **Specificity:** Specificity $= \frac{TN}{TN+FP} = \frac{35}{35+15} = \frac{35}{50} = 0.70$

5. **Finding FN and FP:** Recall $= \frac{TP}{TP+FN} = 0.80$, Total Positives $= TP + FN = 100$ So, $0.80 = \frac{TP}{100} \Rightarrow TP = 80$, therefore $FN = 100 - 80 = 20$ Precision $= \frac{TP}{TP+FP} = 0.75 \Rightarrow 0.75 = \frac{80}{80+FP} \Rightarrow 80 + FP = \frac{80}{0.75} \approx 106.67 \Rightarrow FP \approx 26.67 \approx 27$

6. **Cost of Errors:** Cost of FN = 5 units, Cost of FP = 1 unit Total Cost = (10 FNs × 5) + (20 FPs × 1) = 50 + 20 = 70 units

7. **Prevalence:** Prevalence $= \frac{\text{Number of Positive Instances}}{\text{Total Instances}} = \frac{150}{1000} = 0.15$ or 15%

   This means that 15% of the samples in the dataset belong to the positive class, which is important context for interpreting other metrics. For example, a high Precision value might be easier to achieve when prevalence is low, and the baseline accuracy if we always predicted the negative class would be 85%.

8. **Effect of Threshold:** Increasing the threshold makes the model more "conservative." It will only predict positive if very confident.

   - **Precision** will generally increase (because the model is more sure about its positive predictions)
   - **Recall** will generally decrease (because it will miss more of the actual positives that fall below the higher threshold)

9. **Balanced Accuracy:** Balanced Accuracy $= \frac{\text{Recall}+\text{Specificity}}{2} = \frac{0.9+0.7}{2} = \frac{1.6}{2} = 0.80$

   Why it's more informative for imbalanced data: In imbalanced datasets, standard accuracy can be misleadingly high if the model simply predicts the majority class. For example, if 90% of samples are negative, a model that always predicts negative would have 90% accuracy but would be useless. Balanced Accuracy considers performance on both classes equally, providing a more realistic assessment of model performance across all classes. In this case, the Balanced Accuracy of 0.80 indicates reasonable performance on both positive and negative classes, regardless of their relative frequencies in the dataset.

10. **ROC Point:** The point (FPR, TPR) = (0.1, 0.6) is in the top-left corner of the ROC space, which is good. It means the classifier has a high true positive rate while maintaining a low false positive rate. This is a sign of a good classifier.

11. **AUC Interpretation:** An AUC of 0.5 means the model has no discriminative power. It is as good as random guessing. The ROC curve would be the diagonal line.

12. **Manual ROC Point Calculation:** From Q1: TP=40, FN=10, FP=15, TN=35 TPR (Recall) = $\frac{TP}{TP+FN} = \frac{40}{50} = 0.80$ FPR = $\frac{FP}{FP+TN} = \frac{15}{15+35} = \frac{15}{50} = 0.30$

13. **Probability Ranking:** Sort by probability descending: (0.9, 0.8, 0.6, 0.4, 0.3) with labels (1, 0, 1, 0, 1)

    - **Thresh > 0.9**: Predict 0 for all. TP=0, FP=0, FN=3, TN=2. TPR=0/3=0, FPR=0/2=0. Point (0,0)
    - **Thresh > 0.8**: Predict 1 for first point (0.9), 0 for others. TP=1, FP=0, FN=2, TN=2. TPR=1/3=0.333, FPR=0/2=0. Point (0, 0.333)
    - **Thresh > 0.6**: Predict 1 for first two points (0.9, 0.8). TP=1, FP=1, FN=2, TN=1. TPR=1/3=0.333, FPR=1/2=0.5. Point (0.5, 0.333)
    - **Thresh > 0.4**: Predict 1 for first three points (0.9, 0.8, 0.6). TP=2, FP=1, FN=1, TN=1. TPR=2/3=0.667, FPR=1/2=0.5. Point (0.5, 0.667)
    - **Thresh > 0.3**: Predict 1 for first four points (0.9, 0.8, 0.6, 0.4). TP=2, FP=2, FN=1, TN=0. TPR=2/3=0.667, FPR=2/2=1.0. Point (1.0, 0.667)
    - **Thresh = 0**: Predict 1 for all. TP=3, FP=2, FN=0, TN=0. TPR=1.0, FPR=1.0. Point (1,1)

14. **Weighted Average Precision:** Weighted Avg Precision = $\frac{(80\times0.9)+(20\times0.6)}{80+20} = \frac{72+12}{100} = \frac{84}{100} = 0.84$

15. **Micro vs. Macro Recall: Macro-average Recall:** Simple average = $\frac{0.8+0.5+0.9}{3} = \frac{2.2}{3} \approx 0.733$ **Micro-average Recall:** Assume 100 instances per class:

    - Class 1: TP = Recall × Total = 0.8×100=80, FN=20
    - Class 2: TP=50, FN=50
    - Class 3: TP=90, FN=10
    - Overall TP = 80+50+90=220, Overall FN=20+50+10=80
    - Micro-average Recall = $\frac{220}{220+80} = \frac{220}{300} \approx 0.733$

16. **Threshold for Specificity:** Specificity = $\frac{TN}{TN+FP} \geq 0.95$ We have 100 TN. Let FP be the number of false positives. $\frac{100}{100+FP} \geq 0.95$ $100 \geq 0.95 \times (100+FP)$ $100 \geq 95 + 0.95 \times FP$ $5 \geq 0.95 \times FP$ $FP \leq \frac{5}{0.95} \approx 5.26$ So, the maximum number of FPs we can allow is 5.

17. **Cost-Sensitive Threshold:** The cost of FN (10) is much higher than the cost of FP (1). This means we want to avoid missing positive cases (FN) more than we want to avoid false alarms (FP). To reduce FNs, we need to be more "liberal" in predicting the positive class. This is achieved by **lowering** the decision threshold below 0.5.

18. **PR Curve vs ROC Curve:** In high class imbalance, the number of negatives (TN and FP) is very large. The FPR in the ROC curve can be deceptively optimistic because the denominator (FP+TN) is huge, so even a moderate number of FPs gives a low FPR. The Precision-Recall curve focuses on the performance on the positive class (the minority class of interest) and is not affected by the large number of negatives. It gives a more realistic picture of the model's performance on the important class.

19. **Geometric Mean:** From question 1: Recall = 0.80, Specificity = 0.70 G-Mean = $\sqrt{\text{Recall} \times \text{Specificity}} = \sqrt{0.80 \times 0.70} = \sqrt{0.56} \approx 0.748$

    Difference from F1-score:

    - **F1-score** (calculated in Q3 as 0.762) is the harmonic mean of **Precision and Recall**, focusing only on the positive class performance and balancing false positives vs false negatives.
    - **G-Mean** is the geometric mean of **Recall and Specificity**, considering performance on both positive and negative classes, making it particularly valuable for imbalanced datasets where we want good performance on both majority and minority classes.
    - While F1-score $= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, G-Mean $= \sqrt{\text{Recall} \times \text{Specificity}}$

20. **Precision-Recall Tradeoff:** If we want to increase Recall from 0.6 to 0.8, we need to capture more true positives. This typically means being less strict about predicting the positive class, which will likely result in more false positives. With more false positives, Precision will **decrease**. There is generally a trade-off between Precision and Recall.