Regression Metrics Overview

Regression problems (predicting continuous values) are evaluated using error-based and variance-based metrics. Each metric captures a different notion of model quality, such as average error, squared error, or explained variance.

Mean Absolute Error (MAE)

Formula: $MAE = (1/n) * \Sigma |y_i - \blacksquare_i|$
- Measures average absolute difference between predictions and actual values.
- Intuitive (in same units as target).
- Robust to outliers compared to MSE.
Use case: predicting house prices where absolute deviation matters.

Mean Squared Error (MSE) and RMSE

Formula: $MSE = (1/n) * \Sigma (y_i - \blacksquare_i)^2$
Formula: $RMSE = \mathrm{sqrt}(MSE)$
- Penalizes larger errors more strongly (squares).
- Sensitive to outliers.
Use case: forecasting energy demand where large deviations are costly.

Mean Bias Deviation (MBD)

Formula: $MBD = (1/n) * \Sigma (\blacksquare_i - y_i)$
- Captures average tendency to overpredict (+) or underpredict (-).
- Useful for understanding systematic bias.
Use case: solar power forecasting models often check if system consistently over- or underestimates.

R^2 (Coefficient of Determination)

Formula: $R^2 = 1 - (\Sigma (y_i - \blacksquare_i)^2 / \Sigma (y_i - \blacksquare)^2)$
- Measures fraction of variance in target explained by the model.
- $R^2 = 1$: perfect fit; $R^2 = 0$: no better than mean predictor; $R^2 < 0$: worse than mean.
Use case: general performance measure in regression tasks.

Adjusted R^2

Formula: Adjusted $R^2 = 1 - (1 - R^2) * (n-1)/(n-p-1)$
- Penalizes addition of unnecessary predictors.
- Useful in multiple regression where adding features can inflate plain R^2.
Use case: model selection in econometrics.

Classification Metrics Overview

Classification tasks are evaluated using metrics that capture trade-offs between detecting positives, avoiding false alarms, and balancing both. Key metrics include Precision, Recall, Specificity, F1 Score, ROC, AUC, and PR curves.

Precision and Recall

• Precision = TP / (TP + FP) → Of predicted positives, how many are correct?
• Recall (TPR, Sensitivity) = TP / (TP + FN) → Of actual positives, how many are caught?

Trade-off: lowering threshold increases Recall but decreases Precision, and vice versa.
Use cases:
- High Recall: medical screening (catch every sick patient).
- High Precision: spam filter (avoid blocking legitimate mails).

Specificity and NPV

• Specificity (TNR) = TN / (TN + FP) → Of actual negatives, how many are correctly identified?
• Negative Predictive Value (NPV) = TN / (TN + FN) → Of predicted negatives, how many are correct?

Use cases:
- Specificity: critical where false alarms are costly (e.g., alarms in safety systems).
- NPV: medicine, e.g., 'If test is negative, can I trust it?'.

F1 Score

F1 = 2 * (Precision * Recall) / (Precision + Recall). Harmonic mean of Precision and Recall.

- Arithmetic mean hides imbalance; Harmonic mean punishes it.
Example: Precision=1.0, Recall=0.1 → AM=0.55, F1=0.18.
F1 is high only if both Precision and Recall are high.
Use: spam filters, search ranking, medical tests.

Precision-Recall Curve

Plots Precision vs Recall as threshold varies.
- Left side (low threshold): Recall high, Precision low.
- Right side (high threshold): Precision high, Recall low.

Useful for imbalanced data (fraud detection, cancer, spam). Average Precision (AP) = weighted average of precision across recall levels.

ROC Curve

ROC = Receiver Operating Characteristic (WWII radar origin).
Plots TPR = TP/(TP+FN) vs FPR = FP/(FP+TN).
- Low threshold → high TPR and FPR.
- High threshold → low TPR and FPR.

Shape: staircase of vertical (TP gain) and horizontal (FP gain) steps.
- Random classifier = diagonal (TPR=FPR, AUC=0.5).
- Good model = bows above diagonal.
- Bad model = below diagonal (can be inverted).

AUC (Area Under Curve)

AUC = area under ROC.
Interpretation: probability a random Positive ranks higher than a random Negative.
- AUC=1.0 $\rightarrow$ Perfect separation.
- AUC=0.5 $\rightarrow$ Random guessing.
- AUC<0.5 $\rightarrow$ Worse than random (invert predictions).

Useful as a threshold-independent summary for comparing classifiers.