Evaluation Metrics in Machine Learning

1. Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Key Metrics

- **Accuracy**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Precision (Positive Predictive Value)**

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

  → Among all predicted positives, how many are actually positive.

- **Recall (True Positive Rate / Sensitivity)**

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  → Among all actual positives, how many are correctly predicted as positive.

> **Precision–Recall Trade-off:**
> Precision and recall often move in opposite directions depending on the classification threshold.
> Increasing the threshold usually increases precision but decreases recall.

---

2. Precision–Recall (P–R) Curve

- Sort samples by their predicted likelihood of being positive.
- Vary the threshold and compute precision and recall for each point.
- Plot **Precision (y-axis)** vs. **Recall (x-axis)**.

Interpretation:

- Curves closer to the **top-right** indicate better performance.
- The **Area Under the P–R Curve (AUPRC)** summarizes performance.
- When curves intersect, visual comparison may be inconclusive.

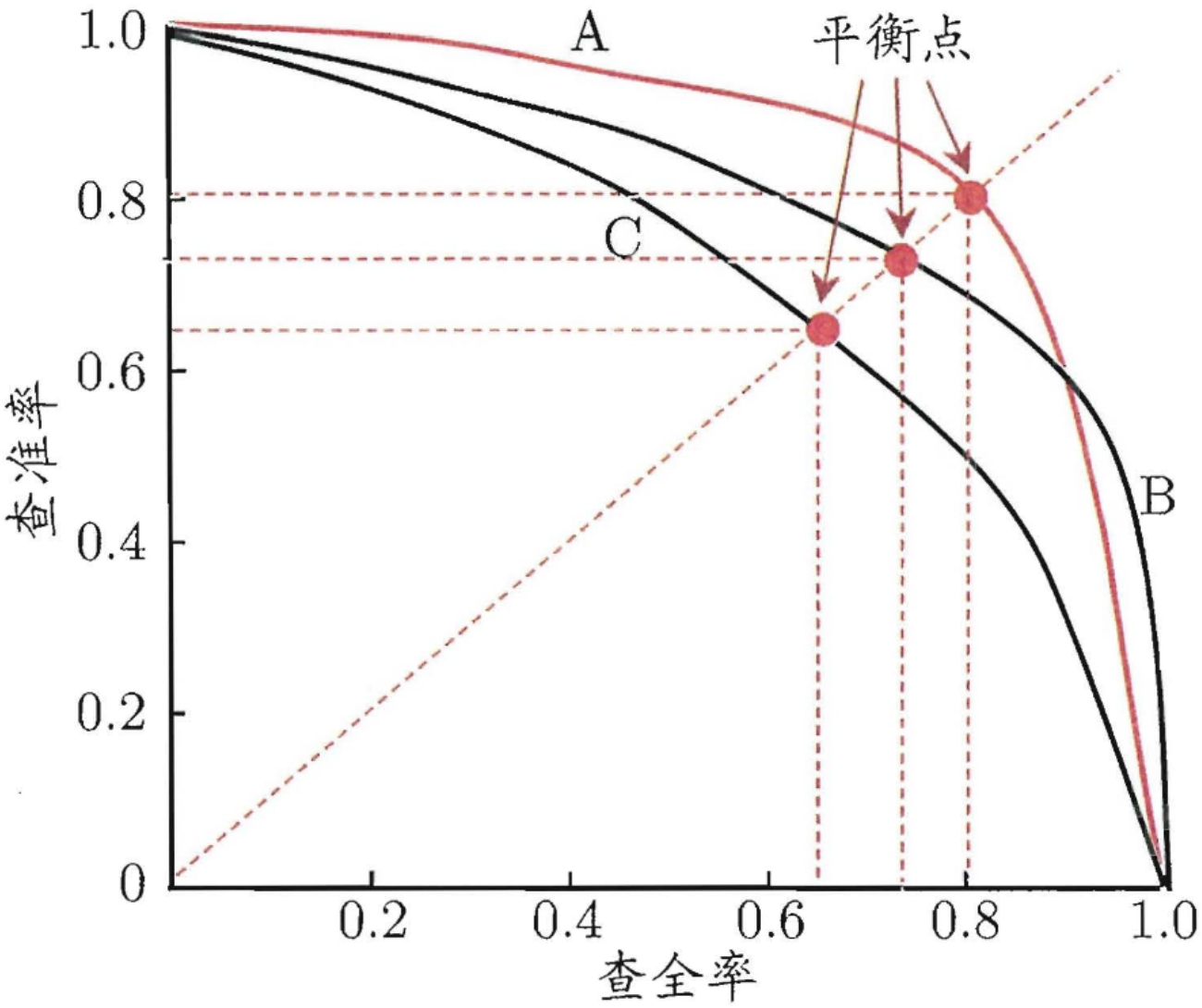


图 2.3 P-R曲线与平衡点示意图

---

3. F-Score and Fβ-Score

The **F1 Score** is the harmonic mean of precision and recall:

$$F_1 = 2 \times \frac{P \times R}{P + R}$$

A generalized form is the **Fβ Score**, which weights recall β times more than precision:

$$F_\beta = (1 + \beta^2) \times \frac{P \times R}{(\beta^2 \times P) + R}$$

- If **β < 1**, precision is more influential.
- If **β > 1**, recall is more influential.
- **β = 1** gives equal weight → standard **F1 Score**.

---

4. ROC Curve and AUC

- **TPR (True Positive Rate)** =

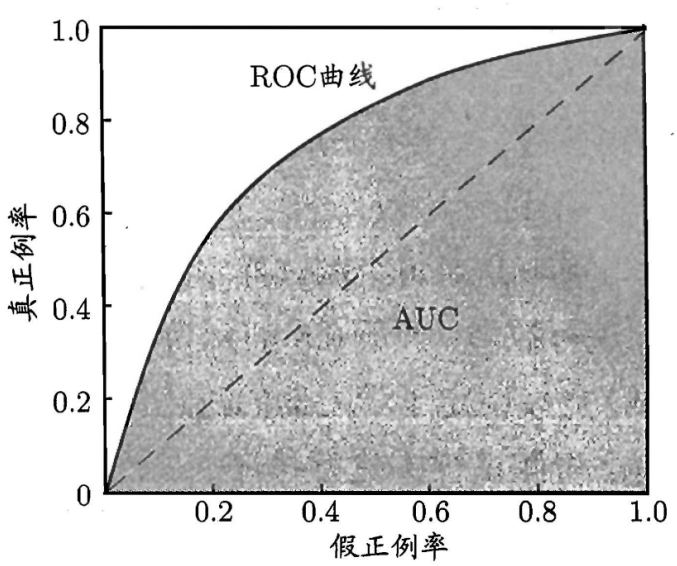
$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

- **FPR (False Positive Rate)** =

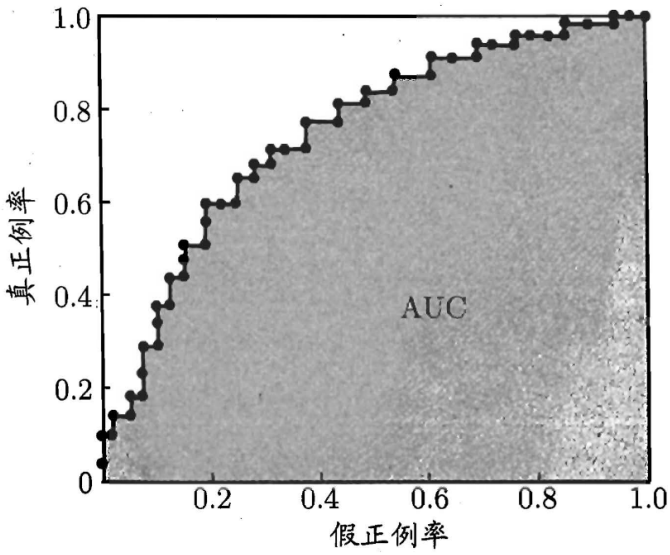
$$\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$$

Plot **TPR (y-axis)** vs. **FPR (x-axis)** as the threshold varies.

- The **diagonal line** (from (0,0) to (1,1)) represents random guessing.
- The **point (0,1)** represents a perfect classifier (100% TPR, 0% FPR).
- A model with an ROC curve entirely above another model's curve performs better.
- The **AUC (Area Under the ROC Curve)** quantifies the overall ranking performance:
  - AUC = 1 → perfect model
  - AUC = 0.5 → random model



(a) ROC 曲线与 AUC



(b) 基于有限样例绘制的 ROC 曲线与 AUC

We can also define the loss by ranking:

形式化地看, AUC 考虑的是样本预测的排序质量, 因此它与排序误差有紧密联系. 给定 m^+ 个正例和 m^- 个反例, 令 D^+ 和 D^- 分别表示正、反例集合, 则排序“损失”(loss)定义为

$$\ell_{rank} = \frac{1}{m^+ m^-} \sum_{\boldsymbol{x}^+ \in D^+} \sum_{\boldsymbol{x}^- \in D^-} \left(\mathbb{I}\left(f(\boldsymbol{x}^+) < f(\boldsymbol{x}^-)\right) + \frac{1}{2}\mathbb{I}\left(f(\boldsymbol{x}^+) = f(\boldsymbol{x}^-)\right) \right), \tag{2.21}$$

即考虑每一对正、反例, 若正例的预测值小于反例, 则记一个“罚分”, 若相等, 则记 0.5 个“罚分”. 容易看出, ℓ_{rank} 对应的是 ROC 曲线之上的面积: 若一个正例在 ROC 曲线上对应标记点的坐标为 (x, y), 则 x 恰是排序在其之前的反例所占的比例, 即假正例率. 因此有

$$\mathrm{AUC} = 1 - \ell_{rank} . \tag{2.22}$$

---

5. Bias and Variance

In supervised learning, the generalization error can be decomposed into:

$$\text{Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Noise}$$

- **Bias:**
  The difference between the average model prediction and the true value.
  High bias → model too simple → underfitting.

- **Variance:**
  Sensitivity of the model to small changes in the training data.
  High variance → model too complex → overfitting.

Relationship

- **High bias, low variance:** underfitting
- **Low bias, high variance:** overfitting
- The goal is to find a balance between the two for best generalization.

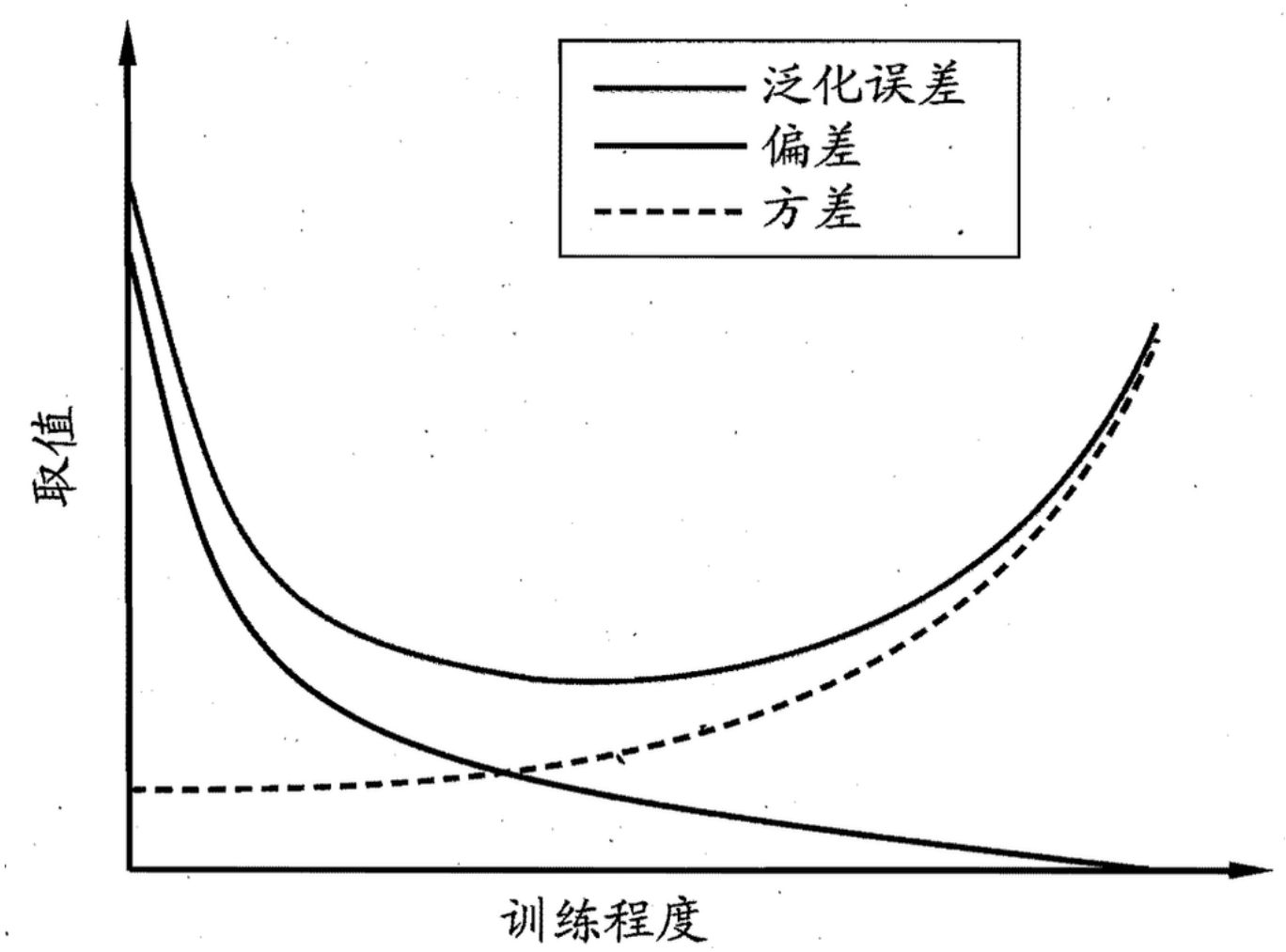


图 2.9 泛化误差与偏差、方差的关系示意图

6. Summary Table

Metric	Formula	Measures	Notes
Accuracy	(TP + TN) / (TP + TN + FP + FN)	Overall correctness	Misleading with imbalanced data
Precision	TP / (TP + FP)	Purity of positive predictions	High precision → few false positives
Recall	TP / (TP + FN)	Coverage of actual positives	High recall → few false negatives
F1	2PR / (P + R)	Harmonic mean of P and R	Balances precision & recall
Fβ	$(1+\beta^2)PR / (\beta^2P + R)$	Weighted F-score	β<1 favors P, β>1 favors R
AUC	Area under ROC curve	Ranking quality	Higher = better
Bias–Variance	—	Generalization behavior	Low bias & variance desired

Questions

Q1. Explain the difference between precision, recall, and accuracy.

- In what types of problems might accuracy be misleading?
- Give an example of a situation where recall is more important than precision.

When is accuracy misleading?

- When classes are imbalanced. Example: 1% positive, 99% negative. A dumb classifier that always predicts “negative” gets 99% accuracy but 0% recall for positives. So accuracy can be high while the model is useless for the minority class.

When is accuracy misleading?

- When classes are imbalanced. Example: 1% positive, 99% negative. A dumb classifier that always predicts “negative” gets 99% accuracy but 0% recall for positives. So accuracy can be high while the model is useless for the minority class.

Q2. How does changing the classification threshold affect precision and recall?

Describe what happens to the Precision–Recall curve as the threshold is varied.

Effect of threshold

- Most probabilistic classifiers output a score s. Prediction = positive if s > t.
- Raising t → fewer positives predicted → precision usually ↑ (fewer false positives) and recall ↓ (more false negatives).
- Lowering t → more positives predicted → recall ↑ and precision ↓.

Precision–Recall (P–R) curve

- Build points by varying t from 1→0 and compute (Recall, Precision) at each point. The curve shows the precision you get for each recall level.
- The area under this curve (AUPRC) summarizes performance; curves closer to top-right are better.

Why might you prefer one threshold over another?

Precision–Recall (P–R) curve

```
•   Build points by varying t from 1→0 and compute (Recall, Precision) at
each point. The curve shows the precision you get for each recall level.
•   The area under this curve (AUPRC) summarizes performance; curves
closer to top-right are better.
•   For imbalanced data, P–R curve is more informative than ROC because it
focuses on performance for the positive (rare) class.
```

Q3. Derive the formula for the Fβ score starting from the harmonic mean definition.

When β < 1, which metric dominates? Why?

How would you interpret β = 2 in a medical diagnosis setting?

• If $\beta < 1$: precision is more influential. Small β shrinks the denominator's β^2 P term, so improvements in precision matter more to F_β.

• If $\beta > 1$: recall is more influential. Large β increases β^2 P in denominator, so having higher recall boosts score more.

Practical interpretation • $\beta = 0.5$ → favor precision (useful when FPs are costly). • $\beta = 2$ → favor recall (useful when FNs are costly).

Q4. Compare and contrast the Precision–Recall (P–R) curve and the ROC curve.

Which one is more informative for imbalanced datasets, and why?

ROC curve (TPR vs FPR):

```
•   Plots TPR = TP/(TP+FN) vs FPR = FP/(FP+TN) across thresholds.
•   Useful when both classes matter and class proportions are not extreme.
•   AUC–ROC measures the model's ability to rank positives above
negatives.
```

Precision–Recall (P–R) curve:

```
•   Plots Precision vs Recall across thresholds.
•   Focuses on the positive class performance directly.
•   When positives are rare, P–R is more informative because precision
directly shows how many retrieved are true positives; ROC can be overly
optimistic because FPR uses TN which dominate when negatives are many.
```

Conclusion

- For imbalanced datasets (rare positives), P–R curve / AUPRC is generally more informative than ROC/AUC.

Q5. What does an AUC value of 0.5 mean?

AUC = 0.5

- Classifier's ranking is equivalent to random guessing (no discriminative power).
- The model cannot rank positives above negatives better than chance.

AUC < 0.5

- Model is worse than random; it systematically ranks negatives above positives.
- Often indicates that the model is inverted (its scores are negatively correlated with true label).
- Quick fix: flip the predicted scores/labels (predict positive when model predicts low score). After flipping, AUC becomes > 0.5.

If inversion is not the issue • Could be due to severe data leakage, label noise, wrong target, or feature processing bug. Investigate data and pipeline.

How to improve AUC generally

- Improve features (feature engineering), handle class imbalance, try different model families, calibrate probabilities, reduce label noise, more training data.

Q6. Explain the bias–variance decomposition of error.

Provide an example of a model or technique that primarily reduces variance.

Example of method that reduces variance

- Bagging / Random Forests: average predictions from many high-variance models (e.g., deep trees on bootstrap samples) reduces variance significantly while keeping bias low. Ensemble averaging reduces the variability of predictions across training sets.

Q7. Suppose two models have nearly identical F1 scores, but one has a higher AUC.

How would you decide which model to deploy?

What additional analysis would you perform?

Considerations

```
1.  Operating point vs ranking: F1 is computed at a particular threshold;
AUC measures ranking across thresholds. If you must choose a single
threshold for deployment, compare precision and recall at that operating
threshold (not only the F1).
2.  Business costs: Which metric aligns with cost? If costs depend on
threshold, choose the model that optimizes the relevant cost function.
3.  Calibration: Check probability calibration (reliability diagrams). A
model with higher AUC but poor calibration may still be worse in practice
if you need well-calibrated probabilities.
4.  Stability and fairness: Compare variance across cross-validation
folds. Higher AUC but high variance might be risky.
5.  PR curves & ROC: Inspect P-R and ROC curves—does the higher AUC model
dominate the other across all recall ranges or only parts?
6.  Precision@k or Recall@k: For ranked retrieval tasks, look at metrics
like Precision@k, Recall@k, NDCG etc.
7.  Latency and resource requirements: If AUC advantage is small but model
B is much slower or heavier, prefer the simpler model.
```

Actionable • Run threshold sweep: compute precision, recall, F1, and business cost at many thresholds. • Use cross-validation to estimate stability. • Check calibration and perform post-hoc calibration (Platt scaling, isotonic) if needed. • If ranking matters (candidate generation), prefer higher AUC; if you must operate at single threshold, prefer the model that gives better metrics at that threshold and exhibits stable behavior.

———

Q8. How can precision, recall, and F1 score be extended to multi-class classification?

Explain macro, micro, and weighted averaging strategies.

When would you prefer each?

Two common multi-class strategies

```
1.  Macro averaging
•   Compute metric per class (treating that class as positive vs all
others), then average across classes:
```

$$\text{Macro-Precision} = \frac{1}{C}\sum_{c=1}^{C} P_c.$$

• Treats each class equally, regardless of frequency. • Good when performance on all classes (including rare ones) matters.

```
2.  Micro averaging
•   Aggregate TP, FP, FN across all classes then compute precision/recall:
```

$$P_{micro} = \frac{\sum_c TP_c}{\sum_c (TP_c + FP_c)}.$$

• Equivalent to computing metric on pooled samples; weights classes by frequency. • Good when overall performance across samples is the goal and class imbalance should reflect importance. 3. Weighted averaging • Weighted average of per-class metrics with class support weights:

$$\text{Weighted-Precision} = \sum_c w_c P_c \text{with} w_c = \text{support}_c/\text{total}.$$

• Compromise: accounts for class frequency while still reporting per-class performance.

Multi-label case

```
•   Metrics computed per label (binary per label) then averaged
(macro/micro/weighted) similar to multi-class.
```

When to prefer each

```
•   Macro: care about rare classes equally (e.g., fairness across
categories).
•   Micro: care about overall sample-level performance (common in multi-
label and imbalanced settings).
•   Weighted: reflect real-world prevalence while still providing per-
class contributions.
```

Q9. If your model's F1 score is low, but AUC is high, what might be happening?

What steps would you take to improve the F1 score without hurting AUC?

What high AUC + low F1 indicates

```
•   Ranking quality is good (model scores separate positives and
negatives), but thresholding / calibration or class imbalance causes poor
precision/recall at chosen threshold.
•   Typical causes:
•   Score distribution leads to poor threshold selection (operating point
is suboptimal).
•   Class imbalance: even with good ranking, threshold 0.5 yields few
```

```
positives.
•   Poor probability calibration: raw scores are not directly usable as
probabilities.
•   Goal mismatch: the metric optimized during training differs from F1 at
the desired threshold.
```

How to improve F1 without hurting AUC

```
1.  Threshold tuning: sweep thresholds on validation set to pick one that
maximizes F1 (or Fβ). AUC unaffected.
2.  Score calibration: apply Platt scaling or isotonic regression to
convert scores to well-calibrated probabilities before thresholding.
3.  Class-weighted loss / focal loss: adjust training to emphasize the
positive class so that the operating point improves.
4.  Post-processing / re-ranking: use a small classifier or rule-based
layer on top of top-ranked candidates to improve precision@k.
5.  Resampling: oversample positives or undersample negatives to get
better decision boundary if model biased due to imbalance.
6.  Use validation metric aligned to F1: train or tune hyperparameters to
maximize F1 (direct optimization or surrogate losses).
```

Key point

```
•   Because AUC measures ranking, you can often get better F1 by choosing
a better threshold or calibrating, not necessarily by changing the model
that already has good AUC.
```

Q10 Theoretically, why is the F1 score a harmonic mean instead of an arithmetic mean of precision and recall?

What advantage does that provide in model evaluation?

Harmonic mean properties

• Harmonic mean of two positive numbers a and b is $\mathrm{H} = \frac{2ab}{a+b}$.

• It is dominated by the smaller of the two values: if one value is very low, the harmonic mean is close to that low value.

• This matches the desired behavior for combining precision and recall: a good combined score should be low if either precision or recall is low.

Why not arithmetic mean?

```
•   Arithmetic mean (P+R)/2 would allow a high value if one metric is high
and the other is low — it doesn’t penalize imbalance strongly.
```

- F1 as harmonic mean penalizes models with imbalanced P and R more heavily: to increase F1, you must raise the lower of P and R.

Mathematical intuition

• Harmonic mean is the reciprocal of the arithmetic mean of reciprocals:

$$\frac{1}{F_1} = \frac{1}{2}\left(\frac{1}{P} + \frac{1}{R}\right)$$

→ Emphasizes that both 1/P and 1/R must be small (i.e., both P and R large).

Practical advantage

- F1 favors balanced performance: it rewards models that perform well on both precision and recall rather than excelling in one and failing in the other. This is often what practitioners want when both FP and FN are costly.

Q11 How does bias–variance trade-off manifest differently in neural networks vs. tree ensembles?

In Tree Ensembles

- Single decision trees have low bias but very high variance (sensitive to data splits).
- Random forests reduce variance via bagging (averaging many decorrelated trees):
- Each tree is trained on bootstrap samples and random feature subsets → lowers variance dramatically while keeping bias low.
- Gradient boosting trees (e.g., XGBoost, LightGBM) work oppositely:
- Sequentially add trees to correct previous errors → reduce bias gradually, but if too many trees, variance increases.
- So, tuning parameters like number of trees, depth, learning rate directly manages the bias–variance balance.

Aspect	Neural Networks	Tree Ensembles
Primary risk	Overfitting (high variance)	Depends: single tree high variance; boosting high variance if deep
Regularization	Dropout, weight decay, early stopping	Number/depth of trees, subsampling, shrinkage
Data need	Require large data to reduce variance	Work well on small/medium data

Q12 How can cross-validation help you estimate bias and variance in practice?

Estimate bias and variance:

```
•   Mean of CV scores ≈ estimate of the model's expected performance (→
reflects bias).
•   Variance of CV scores across folds ≈ estimate of model stability (→
reflects variance).
```

Observation	Implication
Low mean performance + low variance	Model underfits → high bias
High mean performance + high variance	Model overfits → high variance
Moderate mean + low variance	Balanced generalization

Why useful

```
•   Lets you compare models: simpler models → higher bias, lower variance;
complex models → opposite.
•   Helps tune hyperparameters (e.g., regularization, tree depth, number
of layers) by observing how bias–variance pattern changes.
```

Q13 If given only one metric to optimize for a real-world system, which would you choose and why?

Task Type	Typical Primary Metric	Rationale
Classification with balanced classes	Accuracy or AUC	Measures general discrimination
Imbalanced detection (fraud, disease)	F1 or Fβ	Balances precision and recall
Ranking / recommendation	MAP, NDCG, Recall@k	Focuses on top-ranked relevance
Regression	RMSE or MAE	Measures magnitude of prediction error
Generative models	Log-likelihood, Perplexity	Reflects probabilistic fit

2. If I must pick one general metric

I'd pick F1 score (or Fβ) in most real-world classification settings, because:

- It penalizes imbalance between precision and recall.
- It’s threshold-independent (well, threshold-selective but tunable).
- It’s intuitive and interpretable for both engineers and stakeholders.
- You can adjust β to reflect application cost sensitivity (e.g., β>1 for recall-oriented, β<1 for precision-oriented).

However:

- For ranking tasks (like search or recommendation), AUC or NDCG may be superior.
- For continuous regression, RMSE is standard because it directly measures deviation from truth in the same units.