🧠

AI/ML Concept Cards

> ***Simplifying tricky AI/ML concepts into bite-sized cards***

A clear and concise reference to understand the most asked AI/ML fundamentals — from Bias vs Variance to Precision vs Recall, all explained with examples and formulas.

---

1. Bias vs Variance

Bias

- Error due to **overly simplistic assumptions** in the model.
- Leads to **underfitting** → model fails to capture underlying trends.
- Example: Predicting house prices using a straight line when the data is highly non-linear.

Variance

- Error due to **too much sensitivity** to training data.
- Leads to **overfitting** → model memorizes noise, performs poorly on unseen data.
- Example: A decision tree with too many splits that fits training data perfectly but fails on test data.

Bias–Variance Tradeoff

- High Bias = underfitting.
- High Variance = overfitting.
- Goal: Find the right balance for good **generalization**.

---

2. Supervised vs Unsupervised Learning

Supervised Learning

- Data comes with **labels** (input → output).
- Model learns mapping to predict outcomes.
- Examples:
  - Regression → predict house prices.
  - Classification → spam vs non-spam emails.

Unsupervised Learning

- Data has **no labels**.
- Model discovers hidden patterns or clusters.
- Examples:
  - Clustering → customer segmentation.
  - Dimensionality Reduction → PCA to reduce features.

Key Difference

- Supervised = guided by labels.
- Unsupervised = finds patterns without labels.

---

3. Precision vs Recall

Precision (Positive Predictive Value)

- Out of predicted positives, how many are correct?
- Formula: **Precision = TP / (TP + FP)**
- Example: A spam filter marks 100 emails as spam, and 90 are truly spam → Precision = 90%.

Recall (Sensitivity/True Positive Rate)

- Out of actual positives, how many did the model identify?
- Formula: **Recall = TP / (TP + FN)**
- Example: Out of 100 spam emails, the model catches 80 → Recall = 80%.

F1-Score

- Harmonic mean of precision and recall.
- Formula: **F1 = 2 × (Precision × Recall) / (Precision + Recall)**
- Use Case: When you need to balance between precision and recall.

Intuition

- Precision → *How correct are your positive predictions?*
- Recall → *How complete are your positive predictions?*

---

4. Overfitting vs Underfitting

Overfitting

- Model learns both **signal and noise**.
- Very high accuracy on training but poor performance on test data.
- Causes: Too complex model, small dataset, too many epochs.
- Fixes: Regularization, dropout, pruning, more data.
- Example: Polynomial regression curve that passes exactly through every training point.

Underfitting

- Model is too simple → fails to capture data patterns.
- Low accuracy on both training and test data.
- Causes: Too simple model, insufficient features.
- Fixes: More complex algorithm, add features, reduce bias.
- Example: Using linear regression for predicting stock trends.

---

5. Classification vs Regression

Classification

- Predicts **discrete categories**.
- Examples:
    - Fraud detection → Fraud/Not Fraud.

    - Disease diagnosis → Positive/Negative.
- Algorithms: Logistic Regression, SVM, Decision Trees.

Regression

- Predicts **continuous values**.
- Examples:
    - Predicting house prices.
    - Estimating temperature.
- Algorithms: Linear Regression, Ridge, Lasso.

Key Difference

- Classification = labels/categories.
- Regression = numeric/continuous values.

---

6. ROC Curve vs AUC

ROC Curve (Receiver Operating Characteristic)

- Graph of **True Positive Rate (Recall)** vs **False Positive Rate**.
- Shows performance across different thresholds.

AUC (Area Under Curve)

- Numeric measure summarizing ROC curve.
- Range: 0 to 1. Higher = better classifier.

Example

- AUC = 0.95 → excellent classifier.
- AUC = 0.5 → no better than random guess.

---

7. Batch vs Epoch vs Iteration

- **Batch** → Subset of training data processed together.
    - Example: 1,000 samples split into batches of 100 → 10 batches.
- **Epoch** → One complete pass through the dataset.
    - Example: Training on all 1,000 samples once = 1 epoch.

- **Iteration** → One update step (forward + backward pass) for a batch.
    - Example: With 10 batches per epoch → 10 iterations per epoch.

---

8. Generative vs Discriminative Models

Generative Models

- Learn **joint probability distribution P(x,y)**.
- Can generate new data samples.
- Examples: Naive Bayes, GANs.

Discriminative Models

- Learn **conditional probability P(y|x)**.
- Focus on decision boundaries between classes.
- Examples: Logistic Regression, SVM.

Use Cases

- Generative → Data generation, anomaly detection.
- Discriminative → Classification accuracy.

---

9. Parametric vs Non-Parametric Models

Parametric Models

- Fixed number of parameters regardless of dataset size.
- Make assumptions about data distribution.
- Examples: Linear Regression, Logistic Regression.
- Pros: Simple, fast.
- Cons: Less flexible, may underfit.

Non-Parametric Models

- Number of parameters grows with data.
- Make fewer assumptions, more flexible.
- Examples: kNN, Decision Trees.
- Pros: Flexible, adapts to data.

- Cons: Slower, need more data.

---

10. Online vs Batch Learning

Batch Learning

- Train model on entire dataset at once (offline).
- Example: Train on 1M records → deploy model.
- Good for stable data.

Online Learning

- Model updates continuously as new data arrives.
- Example: Real-time recommendation systems that adapt with each user click.
- Good for streaming/real-time data.
