Machine Learning Guide

Contents

1 Classical Machine Learning

1.1 Logistic & Linear Regression

Like Im 5: Linear regression is like drawing a line to guess numbers, like how tall a tree is. Logistic regression guesses if something is one thing or another, like an apple or orange.

Details:

- **Linear Regression**: Predicts numbers using $y = \theta_0 + \theta_1 x$. Cost function: Mean Squared Error (MSE) $J(\theta) = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$.
  - *Closed-form*: $\theta = (X^T X)^{-1} X^T y$.
  - *Gradient Descent*: Update $\theta := \theta - \alpha \frac{\partial J}{\partial \theta}$.
  - *Example*: Predicting house price based on size.
- **Logistic Regression**: Predicts probabilities using sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$. Cost: Log-loss $J(\theta) = -\frac{1}{n}\sum[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$.
  - *Example*: Classifying emails as spam or not.

1.2 SVM (Support Vector Machines)

Like Im 5: SVM builds a big wall to separate things, like apples and oranges, making sure the wall is as wide as possible.

Details:

- **Margin Maximization**: Finds hyperplane maximizing distance to nearest points.
- **Primal/Dual**: Primal optimizes $\min \frac{1}{2}\|w\|^2$, dual uses kernels.
- **Kernel Trick**: Transforms data to higher dimensions for separability.
- **Hinge Loss**: $\max(0, 1 - y_i(w^T x_i + b))$.
- *Example*: Classifying cats vs. dogs based on features.

1.3 Decision Trees, Bagging & Boosting

Like Im 5: Decision trees ask questions to sort things. Bagging asks many friends to vote. Boosting focuses on fixing mistakes.

Details:

- **Decision Trees**: Split data using:
  - *Information Gain*: $IG = \text{Entropy}(\text{parent}) - \sum \frac{n_i}{n}\text{Entropy}(\text{child}_i)$.
  - *Gini Impurity*: $\text{Gini} = 1 - \sum p_i^2$.
- **Bagging**: Random Forest averages multiple trees.
- **Boosting**: AdaBoost weights errors, Gradient Boosting minimizes loss.

- *Example*: Predicting product purchases with Random Forest.

1.4 Loss Functions, Bias & Variance

Like Im 5: Loss functions check how wrong your guess is. Bias is missing the same way every time, variance is guessing all over.

Details:

- **MSE**: $\frac{1}{n}\sum(y_i - \hat{y}_i)^2$.
- **Cross-Entropy**: $-\sum y_i \log(\hat{y}_i)$.
- **Bias-Variance**: Error = Bias^2 + Variance + Irreducible error.
- *Example*: Linear model (high bias) vs. complex tree (high variance).

2 Deep Learning Concepts

2.1 CNNs (Convolutional Neural Networks)

Like Im 5: CNNs look at small parts of a picture, like a cats whiskers, to recognize it.

Details:

- **Layers**: Convolutional (detects features), Pooling (reduces size), Fully Connected (classifies).
- **Regularization**: Dropout (drops neurons), Weight Decay (penalizes large weights).
- **Backpropagation**: Computes gradients via chain rule.
- *Example*: Recognizing digits in MNIST.

2.2 Backpropagation & Optimization

Like Im 5: Backpropagation tells a robot what it did wrong. Optimization helps it learn better.

Details:

- **Backpropagation**: $\frac{\partial L}{\partial w} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w}$.
- **Optimizers**: SGD, Adam (adaptive learning rates).
- **Learning Rate Schedules**: Decay or cosine annealing.
- *Example*: Training a network for stock price prediction.

2.3 RNNs & LSTMs

Like Im 5: RNNs remember words in a story. LSTMs are better at remembering important parts.

Details:

- **RNNs**: $h_t = \tanh(W_h h_{t-1} + W_x x_t)$.
- **LSTMs**: Use forget, input, output gates to manage memory.
- **Gradient Issues**: Vanishing/exploding gradients.
- *Example*: Predicting next word in a sentence.

2.4 Self-Attention & Cross-Attention

Like Im 5: Attention helps focus on important words in a story.

Details:

- **Self-Attention**: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$.
- **Multi-Head**: Multiple attention mechanisms in parallel.
- **Cross-Attention**: One sequence attends to another.
- *Example*: Translating sentences by focusing on key words.

2.5 Transformers

Like Im 5: Transformers read whole sentences to understand or translate them.

Details:

- **Architecture**: Encoder-decoder with attention.
- **Positional Encodings**: Add word order information.
- **Layer Normalization**: Stabilizes training.
- *Example*: BERT answering questions.

2.6 Generative AI (GANs & VAEs)

Like Im 5: GANs draw fake pictures, VAEs imagine new ones.

Details:

- **GANs**: Generator vs. discriminator, loss: $\log D(x) + \log(1 - D(G(z)))$.
- **VAEs**: Maximize variational lower bound with KL divergence.
- *Example*: Generating fake faces or digits.

3 Probability & Statistics

3.1 Random Variables & Distributions

Like Im 5: Rolling a dice gives random numbers. Distributions tell you how likely each number is.

Details:

- **Expectation**: $\mathbb{E}[X] = \sum x_i p(x_i)$.
- **Variance**: $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.
- **Bayes Theorem**: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- *Example*: Predicting rain based on clouds.

3.2 Loss Functions from First Principles

Like Im 5: Loss functions check how wrong your guess is. MLE finds the best guess.

Details:

- **MLE**: Maximize $\prod p(x_i|\theta)$.
- **Regression**: MSE assumes Gaussian errors.
- **Classification**: Cross-entropy for categorical distributions.
- *Example*: Fitting a line to house price data.