🔑 24 Most Important Math Definitions in Data Science

Where Each Math Definition is Used in AI/ML/DL/NN

Optimization & Learning

1. Gradient Descent → Training all models (Linear/Logistic Regression, Neural Networks, SVM, Deep Learning).
2. Lagrange Multiplier → Constrained optimization (SVMs, resource allocation problems).
3. SVM Objective Function → Support Vector Machines for classification.

Probability & Statistics

4. Normal Distribution → Gaussian Naive Bayes, Gaussian Mixture Models, probabilistic inference.
5. Z-Score → Outlier detection, standardization before feeding into ML models.
6. Naive Bayes → Text classification, spam filters, sentiment analysis.
7. MLE (Maximum Likelihood Estimation) → Parameter estimation in probabilistic models (Logistic Regression, GMMs, HMMs).
8. Entropy → Decision Trees (ID3, C4.5), Information Gain.
9. KL Divergence → Variational Autoencoders (VAEs), Reinforcement Learning (policy updates), GAN training.

Linear Models

10. OLS (Ordinary Least Squares) → Linear Regression (closed form).
11. Linear Regression Equation → Regression tasks (forecasting, pricing, trends).
12. MSE (Mean Squared Error) → Loss for regression models.
13. MSE + L2 Reg (Ridge Regression) → Regularized regression to prevent overfitting.
14. Log Loss → Logistic Regression, classification problems.

Neural Networks

15. Sigmoid Function → Logistic Regression, neural network activations (historical, now less common).
16. ReLU → Deep Neural Networks (default activation in CNNs, MLPs).
17. Softmax → Output layer for multi-class classification.
18. Gradient Descent → Backpropagation in NN training.

Similarity & Distance

19. Correlation → Feature selection, exploratory data analysis.
20. Cosine Similarity → Text similarity (NLP), recommendation engines, embeddings (word2vec, BERT).
21. K-Means Objective → Clustering (unsupervised learning).

Model Evaluation

22. F1 Score → Imbalanced classification evaluation (fraud detection, medical diagnosis).
23. R^2 Score → Regression model evaluation (goodness-of-fit).

Linear Algebra / Dimensionality Reduction

24. Eigenvectors/Eigenvalues → Principal Component Analysis (PCA), covariance matrices.
25. SVD (Singular Value Decomposition) → Recommender Systems (collaborative filtering), Latent Semantic Analysis (LSA), dimensionality reduction.

Quick Summary (Cheat-Map)

- Optimization → Gradient Descent, SVM, Lagrange Multiplier
- Probability/Stats → Normal, Z-score, Bayes, MLE, Entropy, KL
- Regression → OLS, Linear Regression, MSE, Ridge, Log Loss
- Neural Networks → Sigmoid, ReLU, Softmax, Gradient Descent
- Similarity/Clustering → Correlation, Cosine, K-Means
- Evaluation Metrics → F1 Score, R^2 Score
- Linear Algebra → Eigenvectors, SVD

1. Gradient Descent

$$\theta_{j+1} = \theta_j - \alpha \nabla J(\theta_j)$$

- θ_j: parameter at step j
- α: learning rate (step size)
- $J(\theta)$: cost/loss function
- $\nabla J(\theta)$: gradient (slope vector)

👉 Used for optimizing ML models

👉 Used in: training ML/DL models (NNs, logistic regression, SVM).

---

2. Normal Distribution

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- x: random variable
- μ: mean
- σ^2: variance
- σ: standard deviation

👉 Used in: probabilistic models, Gaussian NB, anomaly detection.

---

3. Z-Score

$$z = \frac{x - \mu}{\sigma}$$

- x: data point
- μ: mean
- σ: standard deviation

👉 Used in: feature scaling, anomaly detection.

---

4. Sigmoid Function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- x: input value
- Output: squashed between $[0, 1]$.

👉 Used in: logistic regression, binary classification, NN activation.

---

5. Correlation

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{Std}(X) \cdot \text{Std}(Y)}$$

- $\text{Cov}(X, Y)$: covariance between X and Y
- $\text{Std}(X)$: standard deviation of X
- $\text{Std}(Y)$: standard deviation of Y

👉 Strength of linear relationship. Used in: feature selection, EDA.

---

6. Cosine Similarity

$$\text{sim}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

- $A \cdot B$: dot product
- $\|A\|$, $\|B\|$: vector magnitudes

👉 Used in NLP & recommendation.

---

7. Naive Bayes

$$P(y|x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

- $P(y)$: prior probability of class
- $P(x_i|y)$: likelihood of feature given class
- $P(y|x_1 \ldots x_n)$: posterior probability

👉 Simple probabilistic classifier. Used in: text classification, spam filtering.

---

8. Maximum Likelihood Estimation (MLE)

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} P(x_i|\theta)$$

- θ: parameters
- x_i: data samples
- Product = likelihood

👉 Finds best parameter estimates. Used in: estimating parameters in regression, GMMs.

---

9. Ordinary Least Squares (OLS)

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- X: feature matrix
- y: target values
- β: coefficients

👉 Closed-form solution for linear regression. Used in: linear regression.

---

10. F1 Score

$$F1 = \frac{2PR}{P + R}$$

- $P = \frac{TP}{TP+FP}$ (precision)
- $R = \frac{TP}{TP+FN}$ (recall)
  - TP: true positives
  - FP: false positives
  - FN: false negatives

👉 Balance between precision & recall. Used in: classification evaluation (imbalanced data).

---

11. ReLU (Rectified Linear Unit)

$$f(x) = \max(0, x)$$

- x: input
- Output: 0 if negative, else x.

👉 Common neural net activation. Used in: deep neural networks

---

12. R² Score

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- y_i: actual value
- $\hat{y}_i$: predicted value
- $\bar{y}$: mean of actual values

👉 Goodness-of-fit for regression. Used in: regression fit quality.

---

13. Softmax

$$P(y = j|x) = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}}$$

- x_j: score for class j
- K: total classes

👉 Turns logits into probability distribution. Used in: multi-class classification output.

---

14. Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- y_i: actual value
- $\hat{y}_i$: predicted value
- n: number of samples

👉 Loss function for regression.

---

15. MSE + L2 Regularization

$$MSE_{reg} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 + \lambda\sum_{j}\beta_j^2$$

- λ: penalty strength
- β_j: model coefficients

👉 Ridge regression, prevents overfitting.

---

16. Eigenvectors

$$Av = \lambda v$$

- A: square matrix
- v: eigenvector
- λ: eigenvalue

👉 Used in: PCA, covariance analysis.

17. Entropy

$$H(X) = -\sum_{i} p(x_i) \log p(x_i)$$

- $p(x_i)$: probability of outcome x_i

👉 Measure of uncertainty. Used in: decision trees

---

18. K-Means Clustering

$$\min \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

- k: number of clusters
- μ_i: cluster centroid
- S_i: points in cluster i

👉 Groups data into clusters.

---

19. KL Divergence

$$D_{KL}(P||Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

- $P(x)$: true distribution
- $Q(x)$: approximated distribution

👉 Measures how one distribution differs from another. Used in: VAEs, GANs, RL.

---

20. Log Loss

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

- $y_i \in \{0, 1\}$: true label
- $\hat{y}_i \in (0, 1)$: predicted probability

👉 Core classification loss.

---

21. Support Vector Machine (SVM)

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(w \cdot x_i - b))$$

- w: weights
- b: bias
- C: regularization
- y_i: labels (+1/-1)
- x_i: inputs

👉 Finds maximum-margin separating hyperplane. Used in: classification (SVM machines).

---

22. Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

- y: output
- x_j: features
- β_j: coefficients
- ϵ: error

👉 Predicts continuous outcomes.

---

23. Singular Value Decomposition (SVD)

$$A = U\Sigma V^T$$

- A: data matrix
- U, V: orthogonal matrices
- Σ: diagonal matrix of singular values

👉 Used in dimensionality reduction, recommender systems.

---

24. Lagrange Multiplier

$$L(x, \lambda) = f(x) - \lambda g(x)$$

- $f(x)$: function to optimize
- $g(x) = 0$: constraint
- λ: multiplier

👉 Constrained optimization.