Statistical Concepts for Machine Learning

1 Correlation

1.1 Mathematical Definition

Correlation measures the **linear relationship** between two variables, ranging from -1 to $+1$.

Pearson Correlation Coefficient:

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \times \sigma_Y} \tag{1}$$

Where:

- $\text{Cov}(X, Y)$ = covariance between X and Y
- σ_X, σ_Y = standard deviations of X and Y

Alternative formula:

$$r = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

Where:

- x_i = individual values of variable X
- y_i = individual values of variable Y
- $\bar{x}$ = mean of X
- $\bar{y}$ = mean of Y
- n = number of observations

1.2 Interpretation

- $r = +1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship (but nonlinear relationships may exist!)
- $|r| > 0.7$: Strong correlation
- $0.3 < |r| < 0.7$: Moderate correlation
- $|r| < 0.3$: Weak correlation

1.3 Key Insights

- Correlation $\neq$ Causation (classic mistake!)
- Only captures **linear** relationships
- Sensitive to outliers
- Dimensionless (unlike covariance)

2 Covariance

2.1 Mathematical Definition

Covariance measures how two variables **change together**.

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{3}$$

Where:

- $E[\cdot]$ = expected value (mean)
- μ_X = mean of X
- μ_Y = mean of Y

Sample covariance:

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{n - 1} \tag{4}$$

Where:

- x_i, y_i = individual observations
- $\bar{x}$, $\bar{y}$ = sample means
- n = sample size

2.2 Interpretation

- $\text{Cov}(X, Y) > 0$: X and Y tend to increase together
- $\text{Cov}(X, Y) < 0$: When X increases, Y tends to decrease
- $\text{Cov}(X, Y) = 0$: No linear relationship

2.3 Key Insights

- Has units (product of X and Y units)
- Magnitude depends on variable scales
- Used in PCA, portfolio theory, multivariate analysis
- **Covariance Matrix**: Extends to multiple variables

$$\Sigma = \begin{bmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) \end{bmatrix} \tag{5}$$

Diagonal = variances, off-diagonal = covariances

2.4 Relationship Between Covariance and Correlation

$$\text{Correlation} = \text{Normalized Covariance} \tag{6}$$

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \times \sigma_Y} \tag{7}$$

3 Overfitting Detection and Prevention

3.1 What is Overfitting?

Model learns **noise** in training data rather than the underlying pattern. Performs well on training data but poorly on new data.

3.2 Detection Methods

1. Training vs Validation Performance Gap

$$\text{If: Training\_Error} \ll \text{Validation\_Error} \implies \text{Likely overfitting} \tag{8}$$

2. Learning Curves

- Plot error vs training size
- Overfitting: large gap between train/validation curves
- Underfitting: both errors high and close together

3. Cross-Validation Scores

- High variance in CV scores → overfitting
- Use k-fold cross-validation to check consistency

3.3 Prevention Strategies

1. Regularization

- **L1 (Lasso)**: Cost = MSE + $\lambda \sum |w_i|$ → Sparse models
- **L2 (Ridge)**: Cost = MSE + $\lambda \sum w_i^2$ → Shrinks weights
- **Elastic Net**: Combines L1 + L2

Where:

- w_i = model weights/coefficients
- λ = regularization parameter
- MSE = Mean Squared Error

2. More Training Data

- More data = better generalization
- Data augmentation if collection is expensive

3. Reduce Model Complexity

- Fewer features (feature selection)

- Shallower decision trees
- Fewer layers/neurons in neural networks

4. Early Stopping

- Monitor validation error during training
- Stop when validation error starts increasing

5. Dropout (Neural Networks)

- Randomly drop neurons during training
- Forces network to learn robust features

6. Ensemble Methods

- Bagging, Random Forests, Boosting
- Reduces variance through averaging

4 Bias-Variance Tradeoff

4.1 The Fundamental Equation

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error} \tag{9}$$

4.2 Bias

What it is: Error from incorrect assumptions in the model.

- **High Bias** = Underfitting
- Model too simple to capture patterns
- Consistent errors across different datasets
- Example: Linear model for nonlinear data

Reducing Bias:

- Use more complex models
- Add more features
- Remove regularization

4.3 Variance

What it is: Error from sensitivity to training data fluctuations.

- **High Variance** = Overfitting
- Model too flexible, captures noise
- Large changes when trained on different datasets
- Example: Deep decision tree

Reducing Variance:

- Simplify model
- Regularization
- More training data
- Ensemble methods

4.4 The Tradeoff Visualized

Model Complexity Spectrum

- Simple Models
  - High Bias
  - Low Variance
  - Underfitting
- Optimal Complexity
  - Balanced Bias & Variance
  - Minimum Total Error
- Complex Models
  - Low Bias
  - High Variance
  - Overfitting

Sweet Spot: Balance where total error is minimized

4.5 Practical Guidelines

Symptom	**Problem**	**Solution**
High train error, high test error	High Bias	Increase complexity
Low train error, high test error	High Variance	Decrease complexity
Both errors decreasing	Good!	Continue

4.6 Mathematical Derivation (Brief)

For a prediction $\hat{y}$ and true value y:

$$E[(y-\hat{y})^2] = \underbrace{(E[\hat{y}]-y)^2}_{\text{Bias}^2} + \underbrace{E[(\hat{y}-E[\hat{y}])^2]}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Noise}} \tag{10}$$

Where:

- $\hat{y}$ = predicted value
- y = true value

- $E[\cdot]$ = expected value
- σ^2 = irreducible error (noise in data)

Key Insight: You can't simultaneously minimize both bias and variance. Reducing one typically increases the other. The goal is to find the optimal balance for your specific problem.

5 Quick Reference

Correlation: Standardized measure of linear relationship (-1 to $+1$)
Covariance: Unstandardized measure of joint variability (units matter)
Overfitting: Model memorizes training data, fails on new data
Underfitting: Model too simple, misses patterns in training data
Bias: Error from wrong assumptions (underfitting)
Variance: Error from sensitivity to data (overfitting)
Goal: Find model complexity that minimizes total error = Bias^2 + Variance + Noise