Capstone B: Probability & Statistics Foundations

Table of Contents

Linear Algebra in Probability

Linear algebra provides the structure to manage and analyze systems with many random variables. Instead of dealing with individual probabilities, we can work with vectors and matrices that represent the entire state of a system.

- **State Vectors**: A vector `p` can represent a probability distribution, where `p_i` is the probability of the `i`-th state, and `sum(p_i) = 1`.
- **Transition Matrices**: A matrix `T` can represent the evolution of a system, where `T_{ij}` is the probability of transitioning from state `j` to state `i`. The next state is given by `p_{\text{next}} = T\, p_{\text{current}}`.

  Convention note: We adopt the **column-stochastic** convention (columns sum to 1). Some texts use **row-stochastic** matrices (rows sum to 1) and write `p_{\text{next}}^T = p_{\text{current}}^T T`. Both are equivalent when used consistently.

Random Variables and Expectation

- **Random Variable (RV)**: A variable whose value is a numerical outcome of a random phenomenon.
- **Expectation (E[X])**: The long-run average value of a random variable. For a discrete RV `X` that can take values `x_i` with probabilities `p_i`, the expectation is

  $$E[X] = \sum_i x_i\, p_i.$$

  This is simply a **dot product** between the vector of outcomes and the vector of probabilities.
- **Vector of Random Variables**: We can group multiple RVs into a random vector `X = [X_1, X_2, ..., X_n]^T`. The expectation is taken element-wise: `E[X] = [E[X_1], E[X_2], ..., E[X_n]]^T`. ## Covariance Matrix: The Heart of Multivariate Statistics

While variance measures the spread of a single random variable, the **covariance matrix** (Σ or $\mathrm{Cov}(X)$) captures the relationships between multiple random variables in a vector X.

- **Definition**: For a random vector X with mean $\mu = E[X]$, the covariance matrix is an $n \times n$ matrix defined as:

  $$\Sigma = E\left[(X-\mu)(X-\mu)^T\right]$$

- **Entries**:
  - The diagonal entries are the variances: $\Sigma_{ii} = \mathrm{Var}(X_i) = E\left[(X_i - \mu_i)^2\right]$.
  - The off-diagonal entries, Σ_{ij}, are the covariances between variables: $\mathrm{Cov}(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$.

Properties of the Covariance Matrix

1. **Symmetric**: $\Sigma = \Sigma^T$. The covariance between X_i and X_j is the same as between X_j and X_i.
2. **Positive semidefinite**: Σ is always positive semidefinite. This means its eigenvalues are all non-negative ($\lambda_i \geq 0$). This is a crucial property exploited by PCA.

Applications: Markov Chains and Principal Component Analysis (PCA)

Markov Chains

- **Concept**: A mathematical model for a sequence of events where the probability of the next event depends only on the current state.
- **Linear Algebra Connection**:
  - The system is described by a **stochastic matrix** `T` (where each column sums to 1).
  - The state of the system at time `k` is a probability vector `p_k`.
  - The evolution is given by `p_{k+1} = T * p_k`.
  - The **steady-state** or equilibrium distribution of the system is the eigenvector of T corresponding to the eigenvalue $\lambda = 1$. This eigenvector tells us the long-term probabilities of being in each state.

Principal Component Analysis (PCA)

- **Goal**: A dimensionality reduction technique that finds the directions of maximum variance in a dataset.
- **Linear Algebra Connection**:
  1. Start with a data matrix `D`.
  2. Center the data and compute the sample covariance Σ (e.g., $\Sigma = \frac{1}{m} X^T X$ for centered data matrix X).
  3. Perform an **eigen-decomposition** of Σ (which is possible because Σ is symmetric, thanks to the Spectral Theorem).
  4. The **eigenvectors** of Σ are the **principal components** (the new orthogonal axes for the data).
  5. The **eigenvalues** of Σ tell you the amount of variance captured by each principal component.
  6. To reduce dimensionality, you keep the eigenvectors corresponding to the largest eigenvalues.

Bridge to Advanced Topics

- **Gaussian Distributions**: The multivariate normal distribution is completely defined by a mean vector μ and a covariance matrix Σ. The geometry of this distribution (ellipsoidal contours) is determined by the eigenvectors and eigenvalues of Σ.

- **Stochastic Processes**: More advanced processes, like Brownian motion, are studied using the tools of linear algebra combined with calculus.
- **Information Geometry**: This field treats probability distributions as points on a manifold, using tools from differential geometry and linear algebra to define distances and curvature.

Practice Problems

1. A system has two states, A and B. The probability of staying in A is 0.8, and moving from A to B is 0.2. The probability of staying in B is 0.5, and moving from B to A is 0.5. Write the transition matrix `T`.
2. Find the steady-state vector for the Markov chain in problem 1 by finding the eigenvector for $\lambda = 1$.
3. You have a dataset with two variables, `X` and `Y`. `Var(X) = 10`, `Var(Y) = 5`, and `Cov(X, Y) = 3`. Write down the covariance matrix Σ.
4. Why must the eigenvalues of a covariance matrix be non-negative? (Hint: Consider the definition of variance).
5. **Conceptual**: If the first eigenvalue of a covariance matrix is much larger than all the others (e.g., $\lambda_1 \gg \lambda_2, \ \lambda_3, \ \ldots$), what does this imply about the structure of the data?