Assignment 4. Bayesian Networks in Bioinformatics:

Total: 130 points (100 core + 15 conceptual + 15 bonus)

Assignment Overview

This assignment will teach you how to implement **Bayesian Networks** for real-world bioinformatics problems. You'll work with three biological datasets to predict diseases, analyze gene networks, and study protein interactions.

What You'll Learn

By completing this assignment, you will:

- Build Bayesian Networks from biological data
- Predict disease risk using genetic markers
- Analyze gene expression patterns
- Study protein interaction networks
- Apply machine learning to real bioinformatics problems

🗀 Project Files

```
🎁 Your Assignment Files:
├── 📄 src/bayesian_network_project.py     ← MAIN FILE (you complete this)
├── 📄 grade_script.py                     ← Grading script
├── 📄 requirements.txt                    ← Python packages needed
│
💼 data/ (Biological Datasets):
├── 📄 gene_expression.csv                 ← 1200 patients, 50 genes
├── 📄 disease_markers.csv                 ← 800 people, 15 genetic markers
└── 📄 protein_interactions.csv            ← 300 proteins, interaction network
│
💼 src/ (Helper Files - DON'T EDIT):
├── 📄 bayesian_network_model.py           ← Bayesian Network implementation
├── 📄 bioinformatics_utils.py             ← Data loading functions
└── 📄 network_visualizer.py               ← Visualization tools
```

🚀 Getting Started

Step 1: Setup Your Environment

```
# Install required
packages pip install -r
requirements.txt
```

Step 2: Understand the Datasets

Three Datasets:

Each dataset teaches different aspects of Bayesian Networks and reasoning in bioinformatics:

Dataset 1: Gene Expression (Cancer Prediction)

- **1200 patients** (600 healthy, 600 with cancer)
- **50 gene expression levels** per patient (continuous values)
- **Goal**: Predict cancer from gene expression patterns
- **Bayesian Network Type**: Primary disease prediction network

Dataset 2: Disease Markers (Diabetes Prediction)

- **800 people** (400 healthy, 400 with diabetes)
- **15 genetic markers (SNPs)** + clinical variables (age, BMI, etc.)
- **Goal**: Predict diabetes risk from genetic markers
- **Bayesian Network Type**: Secondary risk assessment network

Dataset 3: Protein Interactions

- **300 proteins** involved in cell signaling pathways
- **Interaction scores** between proteins (network data)
- **Goal**: Analyze protein network structure and identify hubs
- **Bayesian Network Type**: Network analysis and graph theory

All three datasets are used in the main script (`src/bayesian_network_project.py`) for different types of Bayesian reasoning!

Step 3: Complete the Assignment

Open `src/bayesian_network_project.py` and complete these sections:

Core Requirements (100 points)

Detailed Point Breakdown:

- [10 pts] STEP 1: Load and Explore Bioinformatics Datasets
- [15 pts] STEP 2: Data Preprocessing and Feature Engineering
- [18 pts] STEP 3: Bayesian Network Structure Learning
- [15 pts] STEP 4: Conditional Probability Calculations
- [18 pts] STEP 5: Probabilistic Inference
- [10 pts] STEP 6: Network Analysis and Visualization
- [10 pts] STEP 7: Protein Interaction Network Analysis
- [4 pts] STEP 8: Model Evaluation and Biological Interpretation

[10 pts] STEP 1: Load and Explore Bioinformatics Datasets

- Load and explore the gene expression, disease markers, and protein interaction datasets
- Calculate basic statistics and correlations

[15 pts] STEP 2: Data Preprocessing and Feature Engineering

- Normalize gene expression data
- Create binary features and SNP interaction features

[20 pts] STEP 3: Bayesian Network Structure Learning

- Learn network structure using correlation-based approach
- Build gene and disease marker networks

[15 pts] STEP 4: Conditional Probability Calculations

- Calculate conditional probabilities for gene and disease marker features

[20 pts] STEP 5: Probabilistic Inference

- Implement Naive Bayes inference for disease prediction
- Evaluate prediction accuracy

[10 pts] STEP 6: Network Analysis and Visualization

- Analyze network properties (density, degree, clustering, etc.)

[10 pts] STEP 7: Protein Interaction Network Analysis

- Build and analyze the protein interaction network
- Identify hub proteins and interaction types

[4 pts] STEP 8: Model Evaluation and Biological Interpretation

- Evaluate model performance using confusion matrix and classification report
- Calculate biological metrics (sensitivity, specificity, precision)

Conceptual Questions (15 pts.)

Answer these questions in your code comments:

1. **Q1 (5 pts)**: Why are Bayesian Networks useful for bioinformatics?
2. **Q2 (5 pts)**: What do gene correlations tell us about disease?
3. **Q3 (5 pts)**: How do protein network hubs affect cell function?

BONUS: Advanced Bayesian Network Analysis (15 pts.)

Complete advanced network analysis:

- Community detection
- Network robustness
- Hub protein identification

📊 Total Points Summary

- **Core Requirements**: 100 points (Steps 1-8)
- **Conceptual Questions**: 15 points (3 questions × 5 pts each)
- **Bonus Analysis**: 15 points (Advanced network analysis)
- **TOTAL**: 130 points

📊 Visualization and Plotting

The project includes comprehensive visualization capabilities:

🎨 Automatic Plot Saving

- All plots are automatically saved to the `plots/` folder
- Plots include timestamps to avoid overwriting
- High-resolution PNG format (300 DPI)

📈 Available Visualization Functions

- `visualize_network()`: Network structure visualization
- `plot_network_metrics()`: Comprehensive network analysis plots
- `plot_community_analysis()`: Community detection visualization
- `plot_correlation_heatmap()`: Feature correlation analysis
- `plot_feature_importance()`: Feature importance ranking
- `plot_performance_metrics()`: Model performance evaluation

Plot Organization

Plots are saved with descriptive filenames:

- `network_[Title]_[Timestamp].png`
- `metrics_[Title]_[Timestamp].png`
- `community_[Title]_[Timestamp].png`
- `correlation_[Title]_[Timestamp].png`
- `importance_[Title]_[Timestamp].png`
- `performance_[Title]_[Timestamp].png`

Tips for Success

Understanding the Code

- Read the helper functions in bioinformatics_utils.py
- Study the Bayesian Network implementation in bayesian_network_model.py
- Use the visualization tools in network_visualizer.py

Biological Context

- **Genes**: Control cell functions and can cause disease when mutated
- **SNPs**: Single genetic variations that affect disease risk
- **Proteins**: Molecular machines that interact to perform cell functions

Code Quality

- Add clear comments explaining your biological reasoning
- Use meaningful variable names
- Handle errors gracefully
- Document your network analysis insights

🆘 Getting Help

Common Issues:

1. **Import Errors**: Make sure you installed all requirements
2. **Data Loading**: Check file paths in the data/ folder
3. **Network Analysis**: Start with simple correlations before complex analysis

Debugging Tips:

- Print intermediate results to understand data flow
- Use small data subsets for testing
- Check the example outputs in the grading script

📋 Submission Checklist

Before submitting your assignment, make sure you have:

- Completed all 8 core steps in bayesian_network_project.py
- Answered all 3 conceptual questions in code comments
- Implemented bonus advanced analysis functions for extra points
- Documented your biological insights and interpretations
- Used meaningful variable names and clear comments

📋 Submission:

- Comply with the submission instructions in BB.
- Locate all the files in a folder and submit a single .zip file in BB after compressing the folder.