INF-2600 Assignment 3: Analyzing Sensor Data for Weather Prediction

Kristina Kufaas
kku020@uit.no

***Abstract*—This report examines a comprehensive dataset of sensor-derived weather data collected over a specified period, focusing on the application of probabilistic models to predict and analyze atmospheric conditions. The primary objective is to employ advanced statistical techniques and Bayesian inference methods to decode the complex dynamics of weather patterns.**

I. Introduction

This report delves into the construction and application of Bayesian Networks to model and predict weather-related phenomena, segmented into two distinct parts. In Part 1, I developed a Bayesian Network based on a provided precode that represents basic relationships among various meteorological variables, where I needed to implement and compare exact inference methods with approximate inference techniques. Additionally, this part involves the creation of two alternative network structures to explore the impact of different relational constructs on the inference outcomes.
Unlike the first part, Part 2 requires the independent construction of a network structure guided by correlation analysis of a brand new dataset, and also includes the application of both exact and approximate inference methods to validate the network model.

II. Theoretical Background

Bayesian learning according to Håkansson and Hartung is about handling the uncertainty of a prediction. [1] It uses Bayes' Probability Theorem and observations to calculate the probability of an outcome. Bayesian learning includes a Bayesian network classifier that represents probability distributions as a graphical model, namely as a directed acyclic graph (DAG).
The basic components of Bayesian Networks are nodes (representing random variables) and edges (representing conditional dependencies). Each node can represent a simple variable, a set of possible outcomes, or a vector of variables. The relationships between these variables are represented as edges connecting the nodes, where each edge points from a "parent" node to a "child" node, indicating a directional dependency.
For each node, a Conditional Probability Distribution (CPD) table is defined, which quantifies the effects of the parents on the child node. This CPD table details the probability of the child node's state given each possible combination of the parents' states. Specifically, the CPD for a variable A with parent B is given by the formula of conditional probability: $P(A \mid B) = P(A \cap B)/P(B)$, calculating the likelihood of A based on the known probabilities related to B. [2] There are two main types of inference methods used in Bayesian networks: exact inference and approximate inference. Exact inference, such as variable elimination can be computationally intensive but provides precise results[3], while approximate inference methods such as direct sampling, rejection sampling, likelihood weighting, Markov chain simulation, Gibbs sampling can be faster but less accurate.

III. Design and Implementation

We were advised to use pgmpy Python library which is specifically designed for creating, learning, and performing inference in Probabilistic Graphical Models (PGMs). It allows users to work with Bayesian Networks and developed to provide a platform for working with probabilistic models in Python and is useful for statistics, machine learning, and AI for reasoning under uncertainty.[4]

1) Task 1: For Task 1, the objective was to implement and analyze a Bayesian network, using a provided precode as a foundation. This involved comparing various inference methods and exploring alternative network structures. I augmented the analysis by adding several functions performing variable elimination [5] to calculate exact probabilities using different combinations of parent and child nodes. For approximate probability calculations, I employed techniques such as rejection sampling, forward (direct) sampling, likelihood weighting, and Gibbs sampling [6]. They were selected due to their efficiency in handling large datasets and complex networks where exact computation could be infeasible and I used the same 4 techniques across the whole assignment. For testing different hierarchies I only added some adjustments in Bayesian network structure, variable dict and the way I created CPDs.

For Task 2, the challenge was to identify most important features through a correlation matrix and integrate these into a Bayesian network. I included all seven recommended parameters, forming connections based on logical relationships inferred from data correlations. The network's structure was predicated on the assumption that higher temperatures potentially leads to alterations in air pressure and humidity. Humidity is also being affected by wind speed, hypothesizing that various aerodynamic forces influence atmospheric moisture. Wind direction was posited to affect wind speed due to directional shifts impacting airflow dynamics. Humidity, which measures the amount of water vapor in the air, is a determinant of precipitation type and once the type of precipitation is

determined, it naturally influences the intensity or amount of precipitation, which is why precipitation type can logically be a parent node for precipitation intensity.[7] I posited wind direction and air temperature as independent parent nodes by analogy with weather from Part 1, because they are probably most important for navigation, safety, and operational planning. Wind direction and speed are necessary for route optimization and fuel management, while air temperature can be important for potential icing conditions. See figure 1 for full overview.

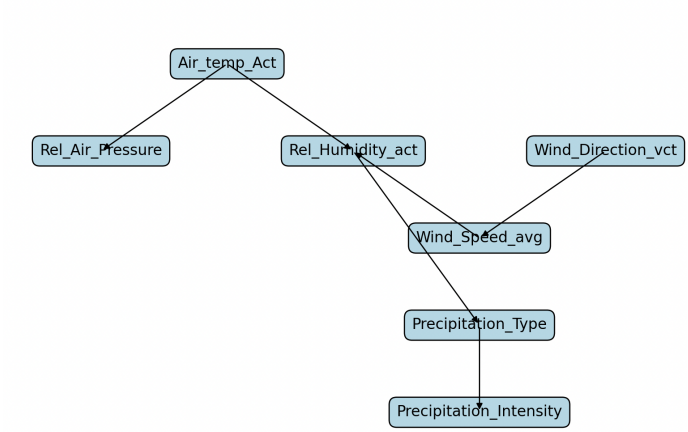

Fig. 1: SPRICE Bayesian network model

IV. Correctness checks and evaluation

A. Correctness checks

I used manual calculation of probability based on conditional probability $P(A \mid B) = P(A \cap B)/P(B)$ for random values to double-check if column/row names are not mixed-up. See `'calculate_probability'` function for reference. In the rest of correctness check I relied on built-in `'check_model'` function.

V. Results

A. Task 1

Among all the combinations evaluated, mid-level precipitation and wind are most commonly recorded during drizzle conditions. It is supported by the highest joint probability observed.

The results for for weather given medium precipitation and specific wind conditions confirm that fog is more likely under low wind conditions and less likely with medium wind, which seems typical atmospheric behavior where fog disperses with increased wind. Drizzle remains likely across both wind conditions. Adding wind as a factor results in a change in the probability distribution of weather conditions under medium precipitation. Specifically, low wind tends to stabilize weather towards foggy conditions, reducing variability, while medium wind increases variability in weather conditions, making more dynamic weather slightly more probable.

1) Exact vs Approximate inference Task 1: **Rejection sampling vs exact inference:**
Probability of high wind when weather is sunny:
0.0824 vs 0.0844
Probability of sunny weather when wind is high:
0.0201 vs 0.0209

Forward sampling vs exact inference: The highest probability, mid-level precipitation/wind and drizzle weather:
0.2970 vs 0.2982

Likelihood weighting vs exact inference:

Weather	Likelihood Weighting	Exact Reference
Fog	0.4593	0.4498
Drizzle	0.4358	0.4508
Rain	0.0570	0.0553
Snow	0.0306	0.0256
Sun	0.0173	0.0185

TABLE I: Conditional probabilities of weather given medium precipitation

Gibbs sampling vs exact inference:

Weather	Low wind	Medium wind
Fog	0.5076	0.4268
Drizzle	0.4590	0.4787
Rain	0.0172	0.0563
Snow	0.0664	0.0170
Sun	0.0098	0.0212

TABLE II: Gibbs sampling, weather given medium precipitation and low or medium wind

Weather	Low wind	Medium wind
Fog	0.5226	0.4180
Drizzle	0.4437	0.4872
Rain	0.0188	0.0564
Snow	0.0664	0.0157
Sun	0.0085	0.0228

TABLE III: Exact inference, weather given medium precipitation and low or medium wind

2) Comparing joint probabilities from all the three-hierarchies using exact inference: For the initial hierarchy and Hierarchy 1 I got equal results. Hierarchy 2 modifies the structure by making Wind directly influence Weather, which then impacts Precipitation. This shift make Hierarchy 2 more weather-centric, where wind plays more important role in determining weather, which then affects precipitation patterns. Under mid conditions for both precipitation and wind, drizzle and fog are more common, with slightly lower probabilities compared to Hierarchy 1. This might indicate that the additional dependency of precipitation on weather (via wind) in Hierarchy 2 dilutes the direct influence of mid-level wind and precipitation on these weather outcomes. Hierarchy 2

shows such entries like low wind leading to fog and snow with high wind, which aren't as prominent in the hierarchies 1 and 2, suggesting that the altered structural dependencies in Hierarchy 2 allow for wind variations to more distinctly influence weather outcomes.

B. Task 2

Among all the combinations evaluated, the combination of very low (under 0,5 m/s) south wind is most commonly recorded. It is supported by the highest (12,5%) joint probability observed.

The results for air temperature given low humidity and specific air pressure show that cold weather (under -10) is more likely (29%) under low air pressure and less likely with high air pressure (12,7%). It can be attributed to the dynamics of rising air, which leads to cooling and often draws in colder air from surrounding areas or altitudes. Under high pressure, the descending air warms and typically brings about more stable and warmer conditions, reducing the likelihood of cold.

There is also a prevalence of mild temperatures (between -10 and 0 degrees) under high air pressure (68%) which might result from the combination of clearer skies, direct sunlight warming the earth more efficiently, and the natural warming of descending air. When there is low pressure (21,7%), with its associated cloud cover and cooler air intrusions, tends to suppress daytime warming, leading to lower temperatures overall.

1) Exact vs Approximate inference Task 2: **Rejection sampling vs exact inference:**

Probability of low wind speed when wind direction is South: 0.1674 vs 0.1667

Probability of South wind direction when wind speed is low: 0.02596 vs 0.2620

Forward sampling vs exact inference: The highest probability, very low south wind: 0.1236 vs 0.1249

Likelihood weighting vs exact inference:

Air temperature	Likelihood Weighting	Exact Reference
Mild	0.6550	0.6643
Warm	0.3041	0.2830
Cold	0.0409	0.0526

TABLE IV: Conditional probabilities of air temperature given high air pressure

Gibbs sampling vs exact inference:

Weather	High pressure	Low pressure
Cold	0.1150	0.2716
Mild	0.6404	0.2214
Warm	0.2046	0.2822
Hot	0.0000	0.2248

TABLE V: Gibbs sampling, air temperature given low humidity and low air pressure

Air temperature	High pressure	Low pressure
Cold	0.1268	0.2886
Mild	0.6818	0.2170
Warm	0.1914	0.2588
Hot	0.0000	0.2355

TABLE VI: Exact inference, air temperature given low humidity and high air pressure

VI. Conclusion

Related to distinctions and similarities between exact and approximate inference methods, while the results from both methods are closely aligned, there is a the subtle differences observed, that might highlight selection of the appropriate inference technique based on the specific requirements.

Exact inference can be especially important in scenarios where even minor inaccuracies can lead to significant consequences, such as in financial forecasting or clinical decision-making processes, while approximate inference maybe offers a more practical solution in situations where computational resources are limited or when a quick estimation is more valuable than absolute precision, for example in weather forecast. Another observation from the iterative execution is the variability in the results with each run. This variability underscores the need for implementing approximate inference techniques over multiple iterations to achieve a more reliable understanding of the outcomes.

I would also like to mention that data set is very special, so it should be a region with predominantly warm temperatures, mostly low wind speeds, and varying air pressure influencing temperature likelihoods — it looks like a region that might have a Mediterranean or subtropical type of climate.

References

[1] A. Håkansson and R. L. Hartung, *Artificial Intelligence. Concepts, Areas, Techniques and Applications*, eng. Studentlitteratur, 2020, ISBN: 9789144125992.

[2] Wikipedia. "Conditional probability." (2024), [Online]. Available: https://en.wikipedia.org/wiki/Conditional_probability (visited on 30/04/2024).

[3] Wikipedia. "Variable elimination." (2024), [Online]. Available: https://en.wikipedia.org/wiki/Variable_elimination (visited on 30/04/2024).

[4] pgmpy. "Reading and writing from pgmpy file formats." (2023), [Online]. Available: https://pgmpy.org/detailed_notebooks/9.%20Reading%20and%20Writing%20from%20pgmpy%20file%20formats.html (visited on 30/04/2024).

[5] pgmpy. "Variable elimination." (2023), [Online]. Available: https://pgmpy.org/exact_infer/ve.html (visited on 30/04/2024).

[6] pgmpy. "Gibbs sampling." (2023), [Online]. Available: https://pgmpy.org/approx_infer/gibbs.html (visited on 30/04/2024).

[7] T. N. C. for Atmospheric Science. “What causes weather?” (2023), [Online]. Available: https://ncas.ac.uk/learn/what-causes-weather/ (visited on 30/04/2024).