iCog Labs

Bayesian Decision Theory for Email Spam Detection

By

Getachew Getu Enyew

January 14, 2025

Abstract

This work explores the application of Bayesian Decision Theory for the classification of email spam using the SMS Spam Collection dataset from UCI. Bayesian classification, a probabilistic approach, integrates prior knowledge and observed data for decision-making. The study outlines the implementation methodology, including the calculation of priors, likelihoods, and posterior probabilities, and demonstrates classification using real-world email data. The classifier achieved an accuracy of 97%, with a precision of 86.42%, recall of 93.96%, and an F1-score of 90.03%, validating its effectiveness. These results demonstrate the classifier's practical utility in accurately identifying spam while minimizing false positives, making it a reliable solution for modern spam detection challenges.

1. Introduction

1.1 Bayesian Probabilistic Theory

A Bayesian network is a representation of probabilistic relationships. These relationships are shown using Directed Acyclic Graphs (DAGs). DAGs are graphs consisting of vertices and directed edges wherein there are no cycles. Each node is a random variable. Referencing Figure 1, the prob ability of a node occurring is the product of the probability that the random variable in the node occurs given that the parents have occurred.

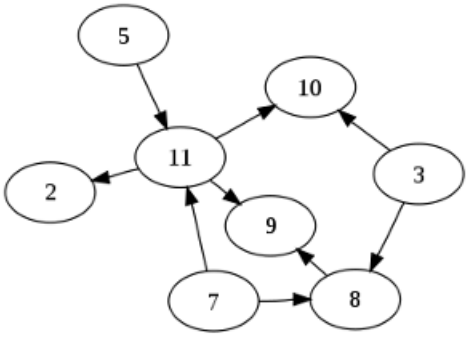

Figure 1: A simple Bayesian Network

Bayesian probabilistic theory is a cornerstone of statistical inference and decision-making, offering a mathematical framework to incorporate prior knowledge with new evidence. Named after Reverend Thomas Bayes, this theory revolves around Bayes' theorem, which provides a systematic approach to updating probabilities based on observed data.

Bayesian probability differs from frequentist interpretations by treating probability as a degree of belief or certainty about an event, rather than simply the frequency of occurrence over repeated trials. This makes Bayesian theory particularly well-suited for applications where prior knowledge, experience, or assumptions play a critical role in the decision-making process.

Bayes' Theorem

The foundation of Bayesian theory is Bayes' theorem, expressed mathematically as:

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

Where:

- $P(H|E)$: Posterior probability - the probability of hypothesis H given evidence E.
- $P(E|H)$: Likelihood - the probability of observing evidence E if hypothesis H is true.
- $P(H)$: Prior probability - the initial belief in hypothesis H, before observing evidence E.
- $P(E)$: Evidence - the total probability of observing E, considering all possible hypotheses.

Bayes' theorem enables the systematic revision of probabilities as new evidence becomes available. This dynamic updating process allows decision-makers to refine their understanding and make better predictions.

Key Components of Bayesian Theory

1. **Prior Probability ($P(H)$):** The prior represents the initial belief about the likelihood of a hypothesis. It is derived from existing knowledge, experience, or assumptions and serves as the starting point for inference. For example, in spam detection, prior probabilities can be based on the historical frequency of spam and non-spam emails in the dataset.
2. **Likelihood ($P(E|H)$):** Likelihood quantifies how well a hypothesis explains the observed evidence. It is calculated based on the conditional probability of the evidence given the hypothesis. In spam detection, this involves assessing how likely certain words or features are to appear in spam versus non-spam emails.
3. **Posterior Probability ($P(H|E)$):** The posterior is the updated probability of the hypothesis after considering the evidence. It combines the prior and likelihood to provide a refined belief about the hypothesis. The posterior serves as the basis for making decisions, such as classifying an email as spam or ham.
4. **Normalization Constant ($P(E|H)$):** The denominator in Bayes' theorem, P(E), ensures that the posterior probabilities sum to 1 across all hypotheses. It is calculated as:

$$\mathrm{P(E)} = \sum_{i=0}^{n} \mathrm{P(E|Hi).\,P(H)}$$

Advantages of Bayesian Theory

- **Incorporates Prior Knowledge:** Bayesian theory allows for the integration of prior beliefs or domain expertise, enabling more informed decision-making in scenarios with limited data.
- **Dynamic Updating:** Probabilities can be updated incrementally as new evidence becomes available, making Bayesian methods ideal for real-time applications.

- **Probabilistic Interpretation:** Bayesian probabilities represent degrees of belief, providing a natural and intuitive way to reason under uncertainty.

1.2 Bayesian Theory in Spam Detection

With the exponential growth of digital communication, email spam poses significant challenges to information security and user experience. Spam emails often include malicious content, fraudulent schemes, or unsolicited advertising, which can compromise privacy, security, and productivity. Identifying spam emails efficiently is essential for maintaining robust communication systems.

Bayesian Decision Theory provides a probabilistic framework for classification problems, making it an ideal choice for spam detection. It combines prior probabilities with observed evidence to compute the posterior probabilities of an email being spam or ham (non-spam). This approach is particularly well-suited for scenarios where prior knowledge and data-driven likelihoods can inform classification decisions.

1.3 Objective

This research implements a Bayesian classifier for email spam detection using the SMS Spam Collection dataset. Key objectives include:

- Defining prior probabilities for spam and non-spam classes.
- Computing word likelihoods based on training data.
- Applying Bayes' theorem to calculate posterior probabilities.
- Classifying emails as spam or ham based on posterior probabilities.
- Evaluating the classifier's performance using precision, recall, F1-score, and accuracy.

2. Methodology

The Bayesian classification process involves several key steps:

2.1 Dataset Description

The SMS Spam Collection dataset, sourced from UCI, contains 5,574 labeled SMS messages. Each message is categorized as either ham (non-spam) or spam. The dataset is preprocessed to tokenize text and convert messages into a format suitable for probabilistic analysis.

2.2 Priors Definition

The priors represent the probability of an email belonging to a specific class (spam or ham) before observing the message content. These are calculated as:

$$P(spam) = \frac{\text{Number of spam messages in training data}}{Total\ messages\ in\ training\ data}$$

$$P(spam) = \frac{\text{Number of ham messages in training data}}{Total\ messages\ in\ training\ data}$$

2.3 Likelihood Calculation

The likelihood represents the probability of specific words appearing in a message given its class. Using Laplace smoothing to handle unseen words, the likelihood for a word www is computed as:

$$P(\text{w}|spam) = \frac{\text{Frequency of w in spam messages} + 1}{\text{Total words in spam messages} + \text{Vocabulary size}}$$

$$P(\text{w}|ham) = \frac{\text{Frequency of w in ham messages} + 1}{\text{Total words in ham messages} + \text{Vocabulary size}}$$

2.4 Posterior Probability Calculation

Bayes' theorem is used to compute the posterior probability of a message being spam or ham:

$$P(\text{spam}|message) = \frac{\text{P(message }|spam)\ .\ P(spam)}{\text{P (message)}}$$

$$P(\text{ham}|message) = \frac{\text{P(message }|\ ham)\ .\ P(ham)}{\text{P (message)}}$$

Since P(message) is constant, it is sufficient to compare the numerators directly:

$$\log \text{Posterior Spam} = \log\big(\text{P(spam)}\big) + \sum_{w \in \text{message}} \log(\text{P(w | spam)})$$

$$\log \text{Posterior ham} = \log\big(\text{P(ham)}\big) + \sum_{w \in \text{message}} \log(\text{P(w | ham)})$$

2.5 Classification Decision

An email is classified as spam if:

$$P(\text{spam}|message) > P(ham|message)$$

Otherwise, it is classified as ham.

2.6 Evaluation Metrics

The classifier's performance is evaluated using:

- **Accuracy**: Percentage of correctly classified messages.
- **Precision**: Percentage of correctly classified spam among predicted spam messages.
- **Recall**: Percentage of actual spam messages correctly identified.
- **F1-Score**: Harmonic mean of precision and recall.

3. Experimental Results

The classifier was evaluated on the SMS Spam Collection dataset, with 80% of the data used for training and 20% for testing. The results are presented below:

Prior Probability for Spam and Ham for spam detection

```
Priors:
P(spam) = 0.1342
P(ham) = 0.8658
```

Likelihood Probability for Spam and Ham spam detection

```
Likelihood of 'squeeeeeze!!' given spam: 0.000038
Likelihood of 'squeeeeeze!!' given ham: 0.000015
Likelihood of 'this' given spam: 0.002528
Likelihood of 'this' given ham: 0.002744
Likelihood of 'is' given spam: 0.004367
Likelihood of 'is' given ham: 0.008308
Likelihood of 'christmas' given spam: 0.000077
Likelihood of 'christmas' given ham: 0.000075
Likelihood of 'hug..' given spam: 0.000038
Likelihood of 'hug..' given ham: 0.000015
Likelihood of 'if' given spam: 0.001149
Likelihood of 'if' given ham: 0.004131
Likelihood of 'u' given spam: 0.003716
Likelihood of 'u' given ham: 0.010500
Likelihood of 'lik' given spam: 0.000038
Likelihood of 'lik' given ham: 0.000075
Likelihood of 'my' given spam: 0.000345
Likelihood of 'my' given ham: 0.008785
Likelihood of 'frndshp' given spam: 0.000038
Likelihood of 'frndshp' given ham: 0.000015
Likelihood of 'den' given spam: 0.000038
Likelihood of 'den' given ham: 0.000358
Likelihood of 'hug' given spam: 0.000038
Likelihood of 'hug' given ham: 0.000030
Likelihood of 'me' given spam: 0.000690
...
```

Table 1: Performance Metrics

Metrics	**Value**
Accuracy	97%
Precision	86.42%
Recall	93.96%
F1-Score	90.03%

Analysis

- The high accuray indicates the classifier effectively minimizes false positives, ensuring legitimate messages are not flagged as spam.
- The high recall demonstrates its ability to identify most spam messages correctly.
- The F1-score highlights a balance between precision and recall, making the classifier reliable for practical use.

4. Discussion

The Bayesian classifier achieved excellent results, with an accuracy of 97%. The use of Laplace smoothing ensured robust handling of unseen words in test data. High precision and recall scores highlight the classifier's practical applicability in real-world scenarios, where misclassification of legitimate messages or undetected spam can have significant consequences.

Key observations:

1. Spam messages frequently include promotional keywords and phrases, which the classifier effectively identified.
2. Rare words in ham messages contributed to occasional misclassifications, suggesting room for further optimization, such as weighting words based on context.

5. Conclusion

This work demonstrates the effectiveness of Bayesian Decision Theory in email spam detection. The classifier's ability to combine prior knowledge with observed data makes it a powerful tool for probabilistic classification tasks. Future work could explore integrating more advanced preprocessing techniques, such as stemming and stop-word removal, to further enhance performance.

6. References

1. Eberhardt, Jeremy J. "Bayesian spam detection." Scholarly Horizons: University of Minnesota, Morris Undergraduate Journal 2.1 (2015): 2.
2. https://en.wikipedia.org/wiki/Bayesian_statistics
3. https://en.wikipedia.org/wiki/Bayes%27_theorem
4. Alpaydin, E. (2016). *Introduction to Machine Learning*. MIT Press.
5. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
6. UCI Machine Learning Repository. SMS Spam Collection Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/sms+spam+collection