# Learning Non-Linear Functions for Text Classification

Cohan Sujay Carlos[1] Geetanjali Rakshit[2]

Aiaioo Labs, Bangalore, India

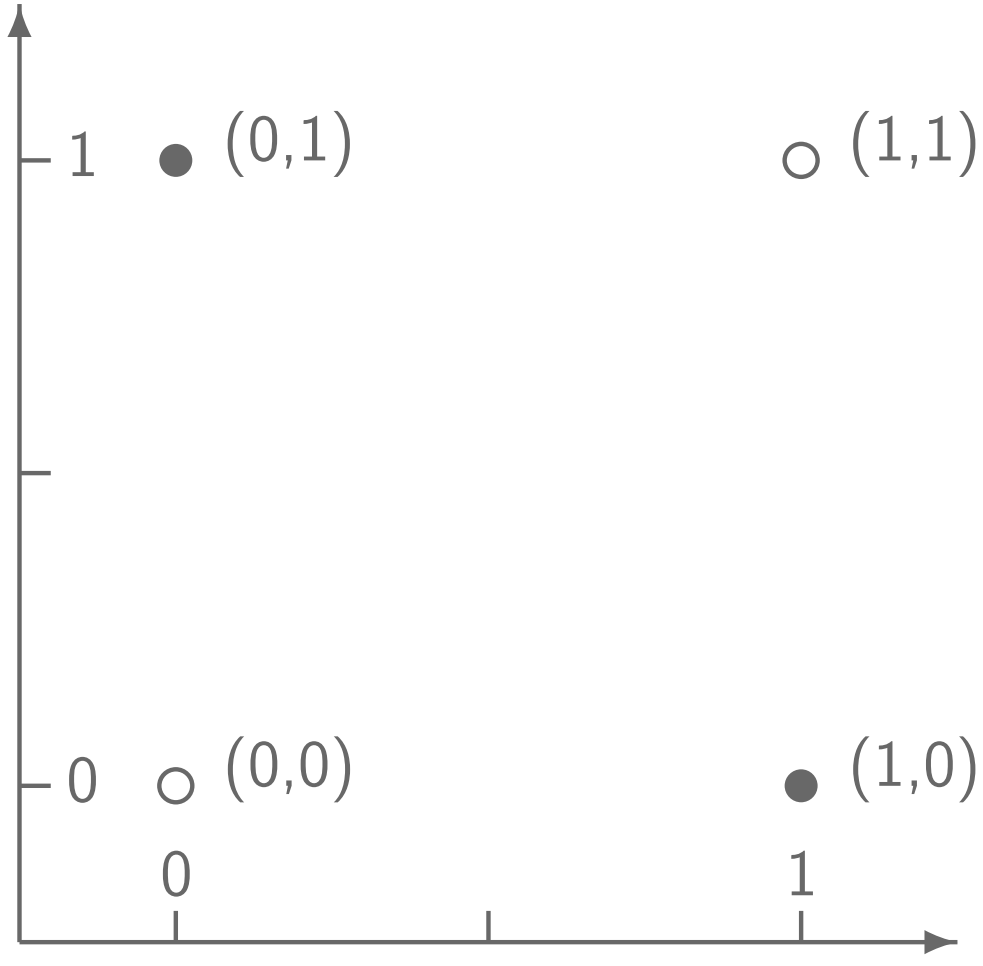
IIT-B Monash Research Academy, Mumbai, India

December 20, 2016

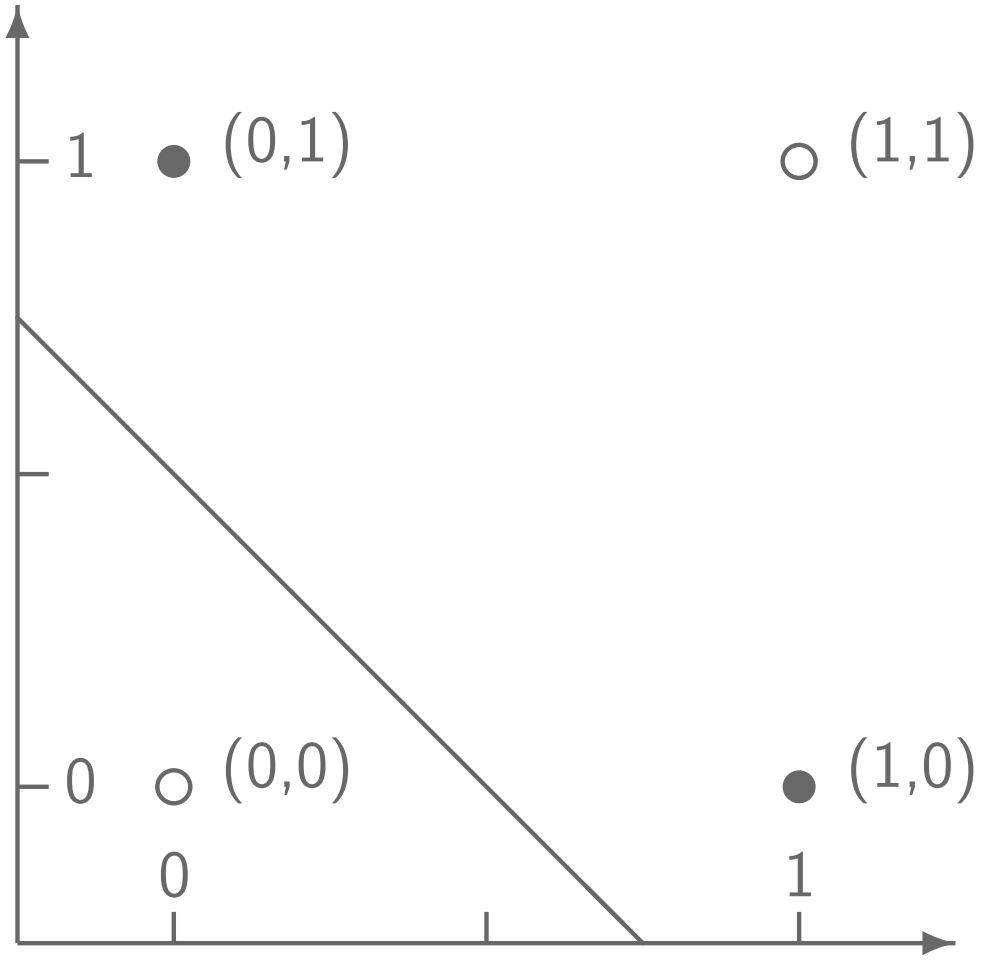
# Outline

# Outline

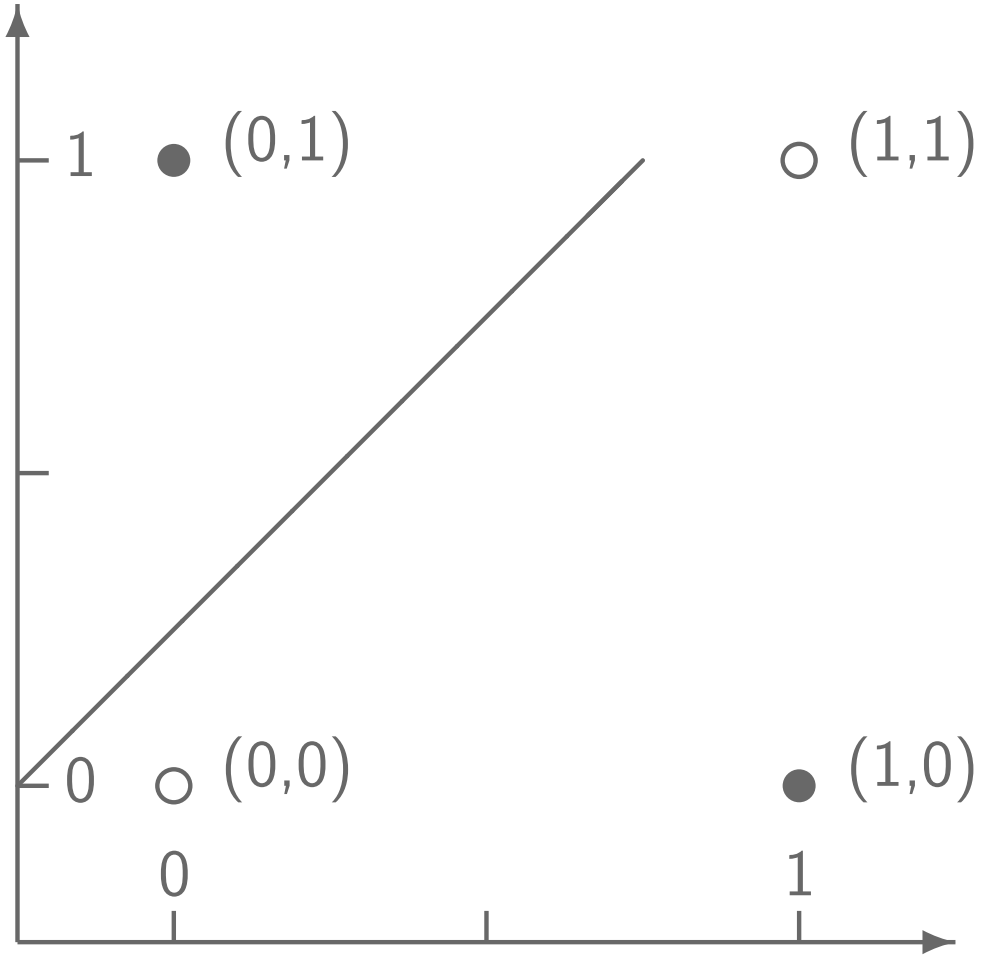
# XOR



# XOR



# XOR



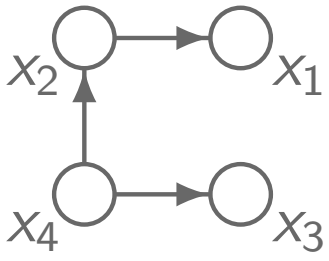
# Motivation

Multilayer neural networks have been shown to be capable of learning non-linear boundaries.

Questions:

- Is there something special about neural networks?
- Can probabilistic classifiers also learn non-linear boundaries?
- In particular, can multilayer Bayesian models do so and under what conditions?

# Bayesian Models = Directed PGMs

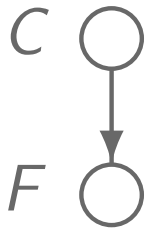



The joint probabilities factorise according to Equation 1.

$$P(x_1, x_2 \ldots x_n) = \left( \prod_{1 \leq i \leq n} P(x_i | x_{a(i)}) \right) \tag{1}$$

where $a(i)$ means *ancestors of node* $i$.

# Naive Bayes




The naive Bayes classifier is a special case two layers deep.

$$P(F, C) = P(F|C)P(C) = \prod_{f \in F} P(f|C)P(C) \qquad (2)$$

The root node represents a class label from set $C = \{c_1, c_2 \ldots c_k\}$ and the leaf nodes the set of features $F = \{f_1, f_2 \ldots f_i\}$.

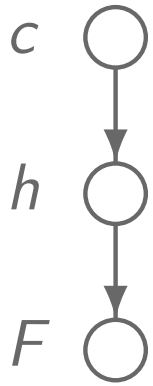
# Outline

# 3-Layer Bayesian Models


In a 3-layer Bayesian model, there is also a set of hidden nodes $H = \{h_1, h_2 \ldots h_j\}$.

$$P(F, H, C) = P(F|H, C)P(H|C)P(C) \tag{3}$$

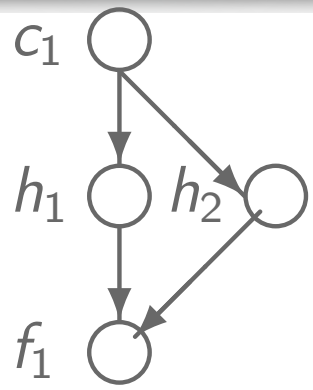
# 3-Layer Bayesian Models




From $P(F, h, c) = P(F|h, c)P(h|c)P(c)$ we obtain 4 and 5.

$$P(F, c) = \sum_h P(F|h, c)P(h|c)P(c) \tag{4}$$

$$P(F|c) = \sum_h P(F|h, c)P(h|c) \tag{5}$$




Starting from $f$ instead of $F$, we get 8.

$$P(f, h, c) = P(f|h, c)P(h|c)P(c) \tag{6}$$

$$P(f, c) = \sum_h P(f|h, c)P(h|c)P(c) \tag{7}$$

$$P(f|c) = \sum_h P(f|h, c)P(h|c) \tag{8}$$

# Product of Sums

You can derive the likelihood as follows.

$$P(F|c) = \prod_{f} P(f|c) \tag{9}$$

Substituting Equation 8 into Equation 9, we get Equation 13.

$$P(F|c) = \prod_{f} \sum_{h} P(f|h,c)P(h|c) \tag{10}$$

# Sum of Products

You can also derive the likelihood as follows.

$$P(F|h,c) = \prod_{f} P(f|h,c) \tag{11}$$

Substituting Equation 11 into Equation 5, we get Equation 14.

$$P(F|c) = \sum_{h} \left( \prod_{f} P(f|h,c) \right) P(h|c) \tag{12}$$

# Two Likelihood Equations

## Product of Sums

$$P(F|c) = \prod_{f} \sum_{h} P(f|h,c)P(h|c) \tag{13}$$

## Sum of Products

$$P(F|c) = \sum_{h} \left( \prod_{f} P(f|h,c) \right) P(h|c) \tag{14}$$

# Outline

# Proof of Linearity

If any classifier can be proved to be equivalent to a multinomial naive Bayes classifier, then it can only learn linear decision boundaries.

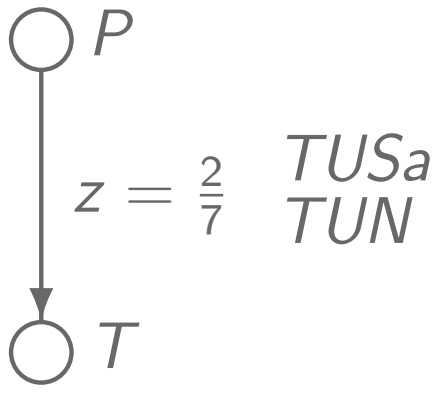
# Proof of Linearity

Consider a toy dataset consisting of two sentences, both belonging to the class Politics (P):

- The United States and (TUSa)
- The United Nations (TUN)

# Proof of Linearity

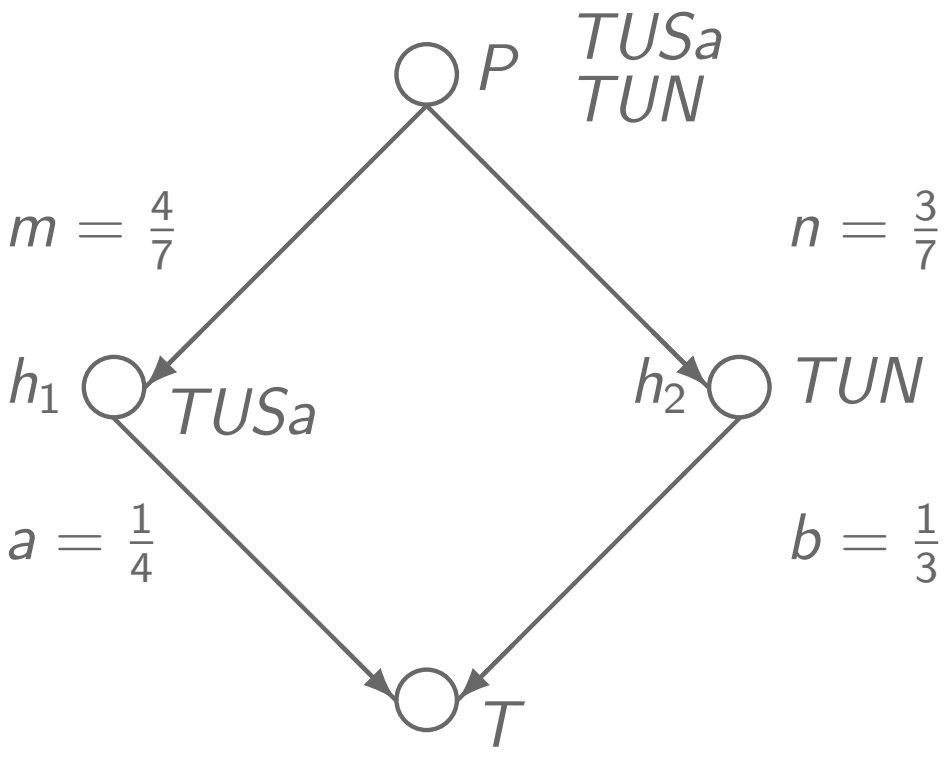
One of the parameters a naive Bayes classifier would learn is $P(The|Politics) = z$.




If a directed PGM learns the same likelihoods, which is a product of these parameters, it can only learn linear decision boundaries.

# Proof of Linearity - Hard EM

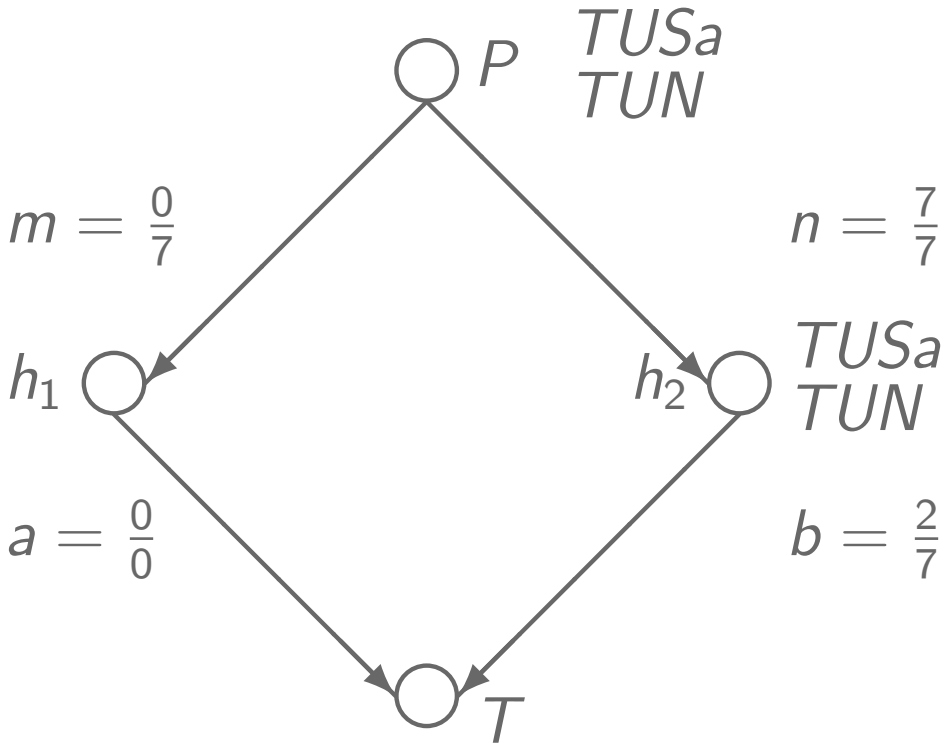
The parameters of a multinomial 3-layer Bayesian classifier:




From Equation 13, $z = am + bn = \frac{1}{4}\frac{4}{7} + \frac{1}{3}\frac{3}{7} = \frac{2}{7}$

# Proof of Linearity - Hard EM

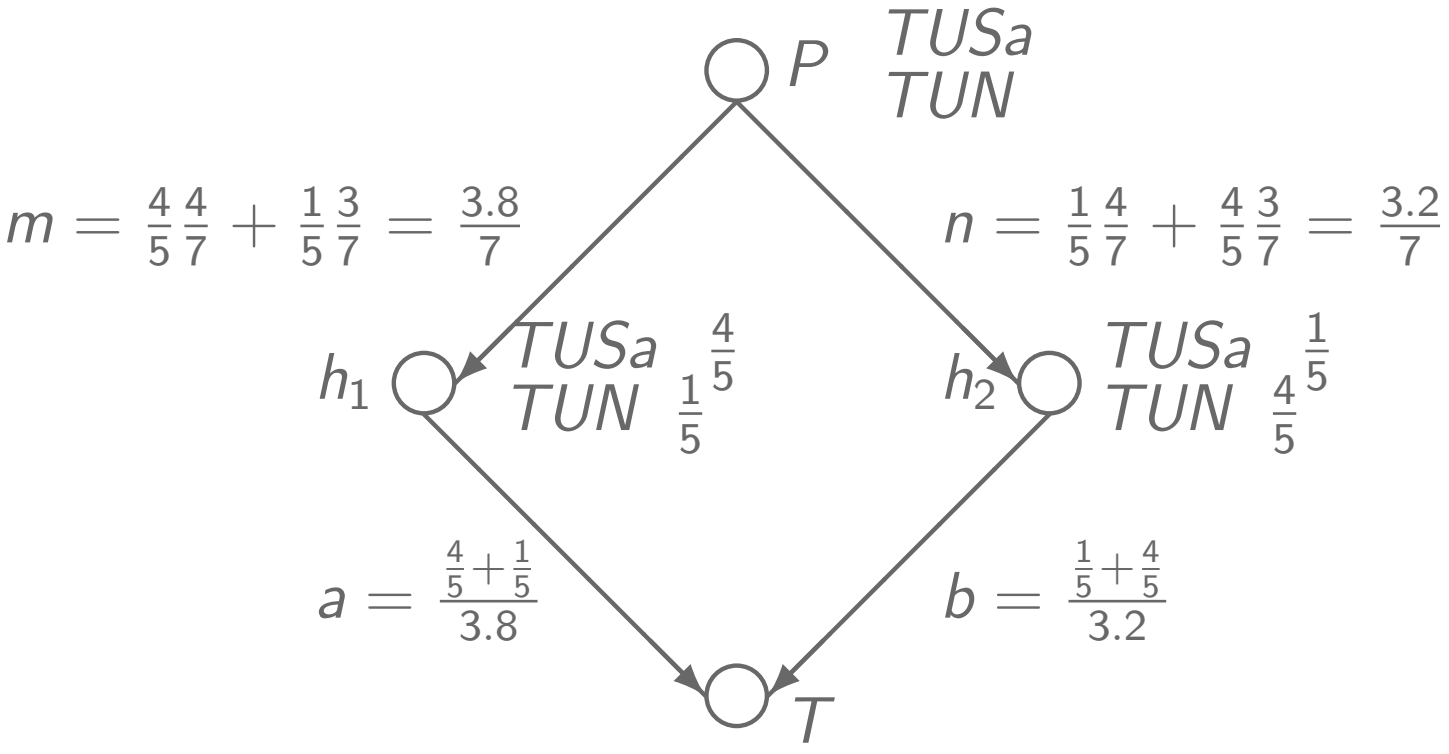
The parameters of a multinomial 3-layer Bayesian classifier:




From Equation 13, $z = am + bn = \frac{0}{0}\frac{0}{7} + \frac{2}{7}\frac{7}{7} = \frac{2}{7}$

# Proof of Linearity - Soft EM

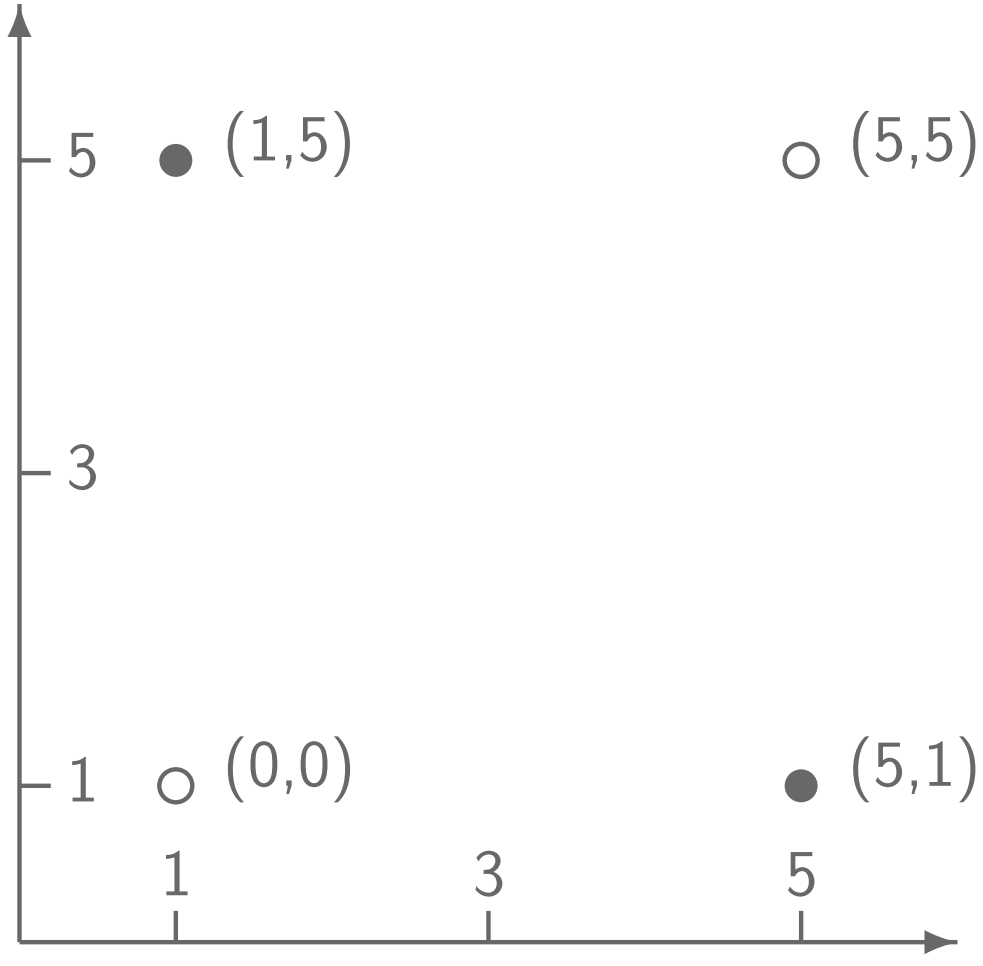
The parameters of a multinomial 3-layer Bayesian classifier:



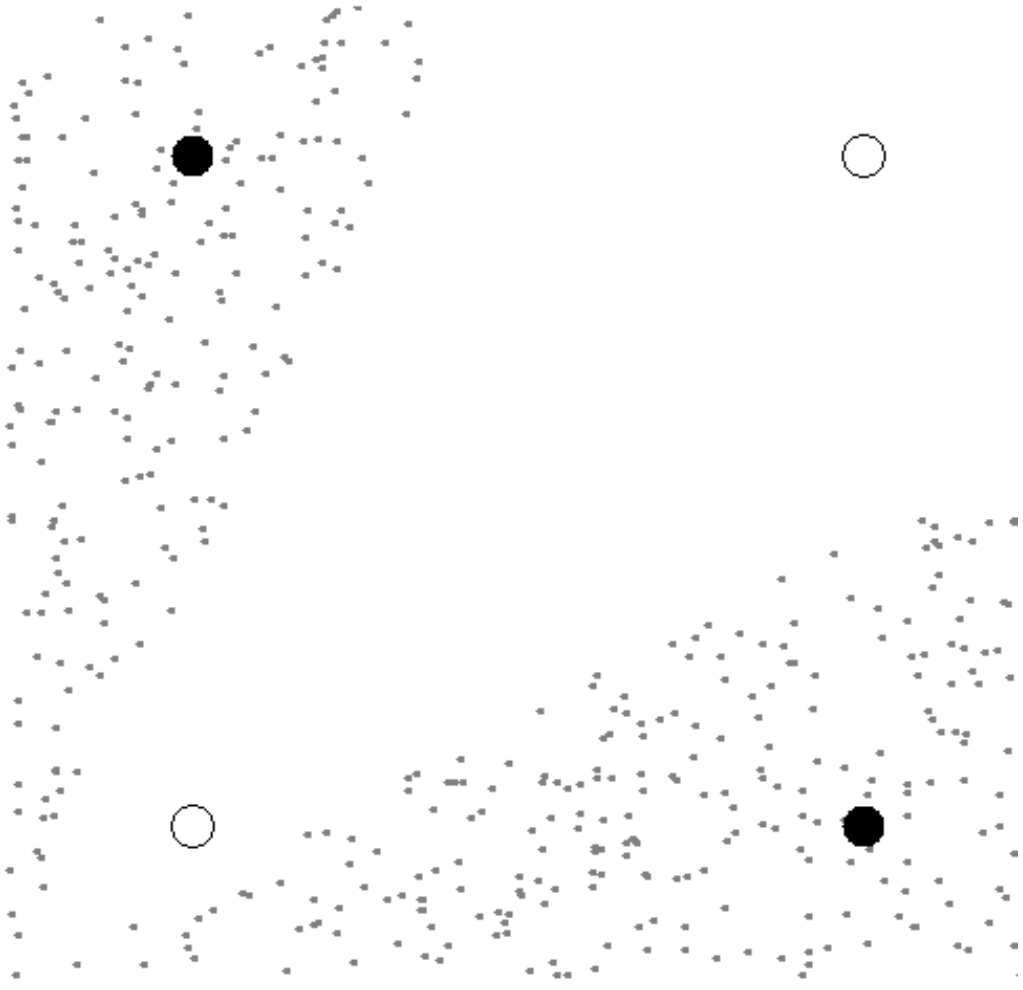

From Equation 13, $z = am + bn = \frac{1}{3.8}\frac{3.8}{7} + \frac{1}{3.2}\frac{3.2}{7} = \frac{2}{7}$

# Outline

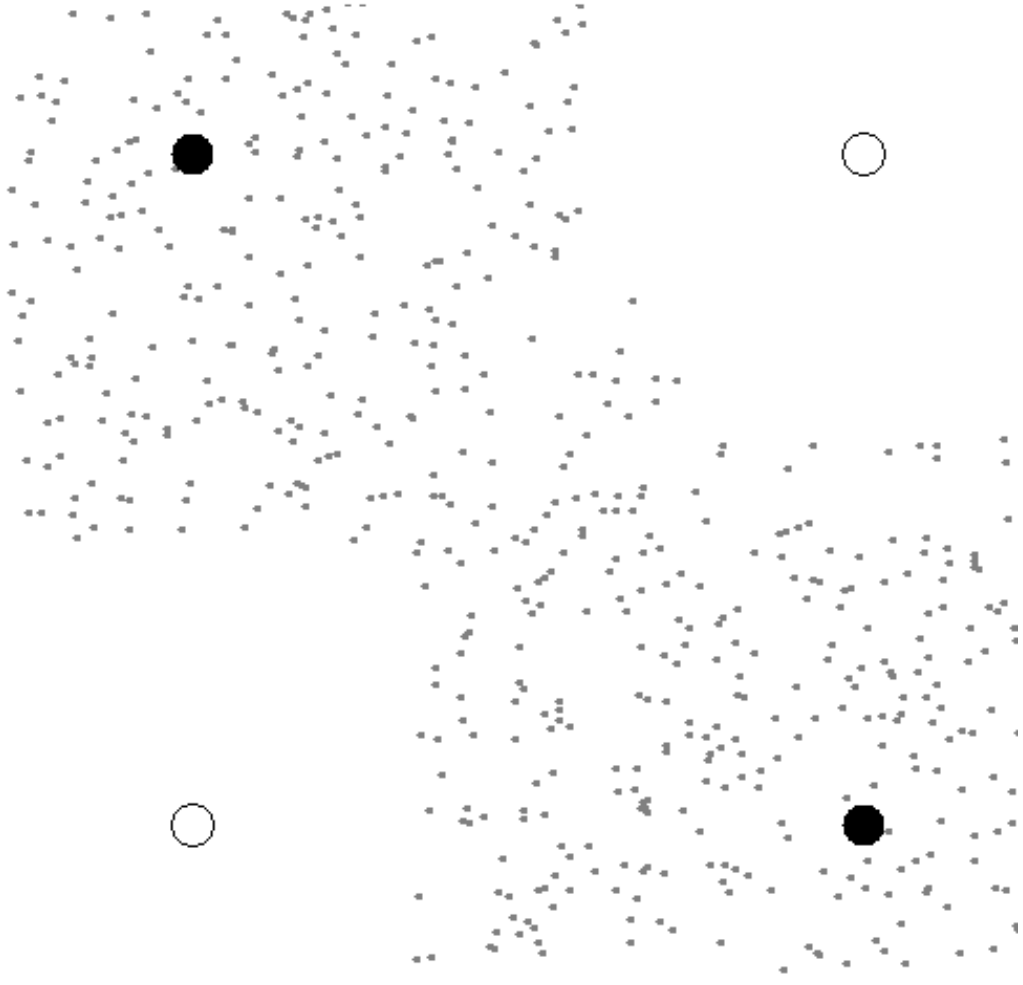
# Shifted and Scaled XOR



# Sum of Products - Multinomial



# Sum of Products - Gaussian



# Outline

# Accuracy on Reuters & 20 Ng

| Classifier | R8 | R52 | 20Ng |
|---|---|---|---|
| Naive Bayes | 0.955 | 0.916 | 0.797 |
| MaxEnt | 0.953 | 0.866 | 0.793 |
| SVM | 0.946 | 0.864 | 0.711 |
| Non-linear 3-Layer Bayes | **0.964** | **0.921** | **0.808** |

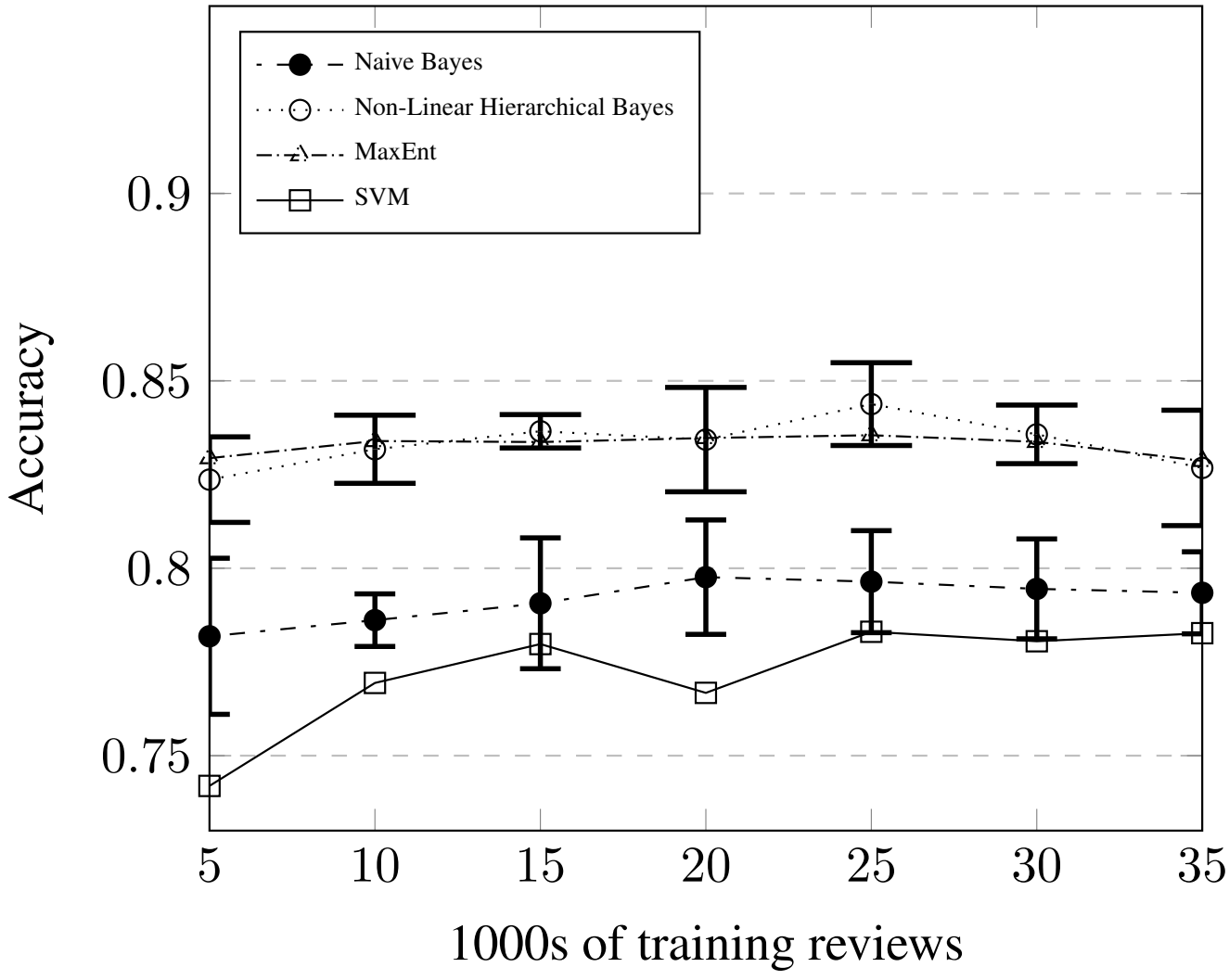
Table: Accuracy on Reuters R8, R52 and 20 Newsgroups datasets.

# Accuracy on Movie Reviews

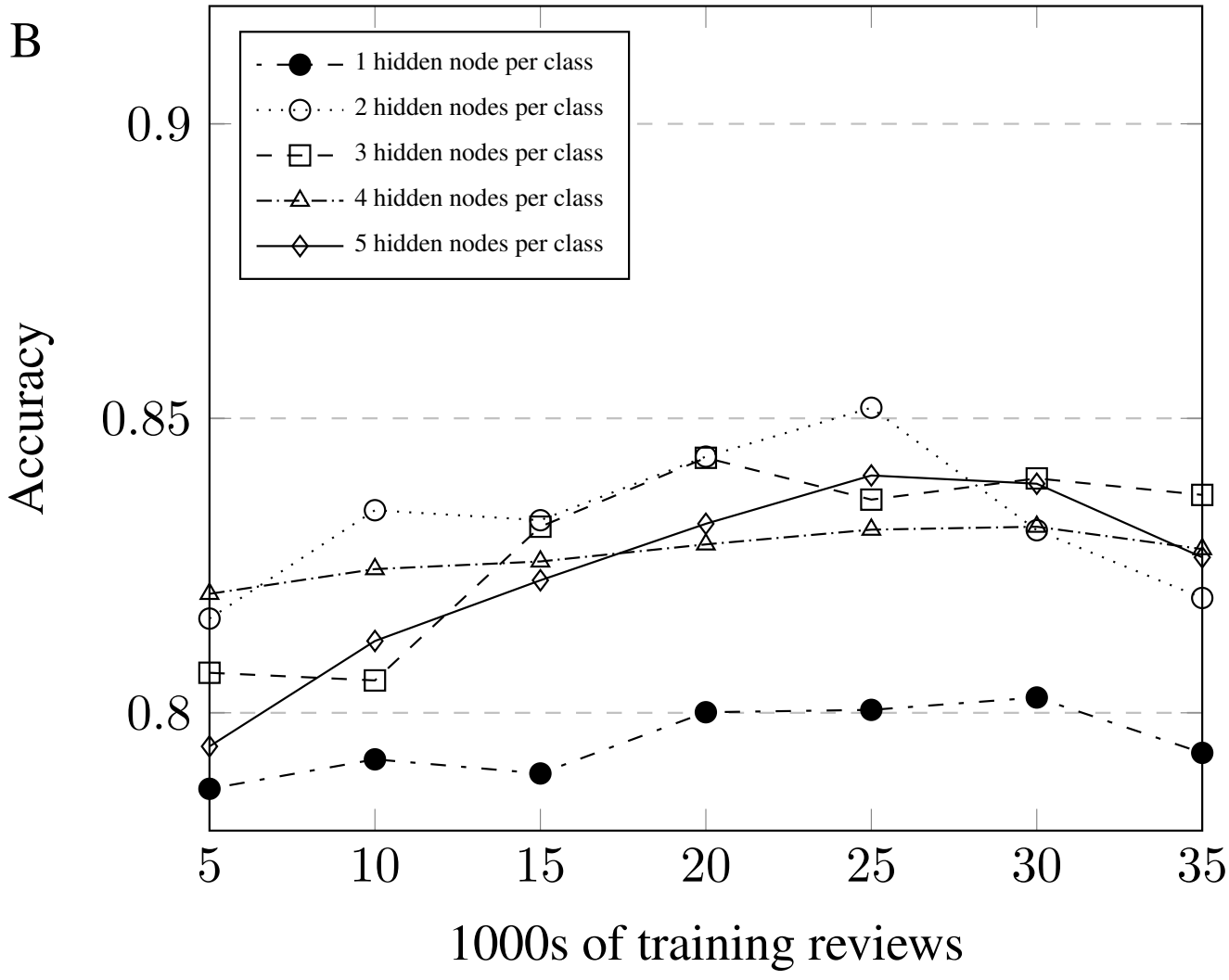
| Classifier | Movie Reviews |
|---|---|
| Naive Bayes | $0.7964 \pm 0.0136$ |
| MaxEnt | $0.8355 \pm 0.0149$ |
| SVM | $0.7830 \pm 0.0128$ |
| Non-linear 3-Layer Bayes | $\mathbf{0.8438} \pm 0.0110$ |

Table: Accuracy on Large Movie Review dataset.

# Accuracy on Movie Reviews



# Accuracy on Movie Reviews

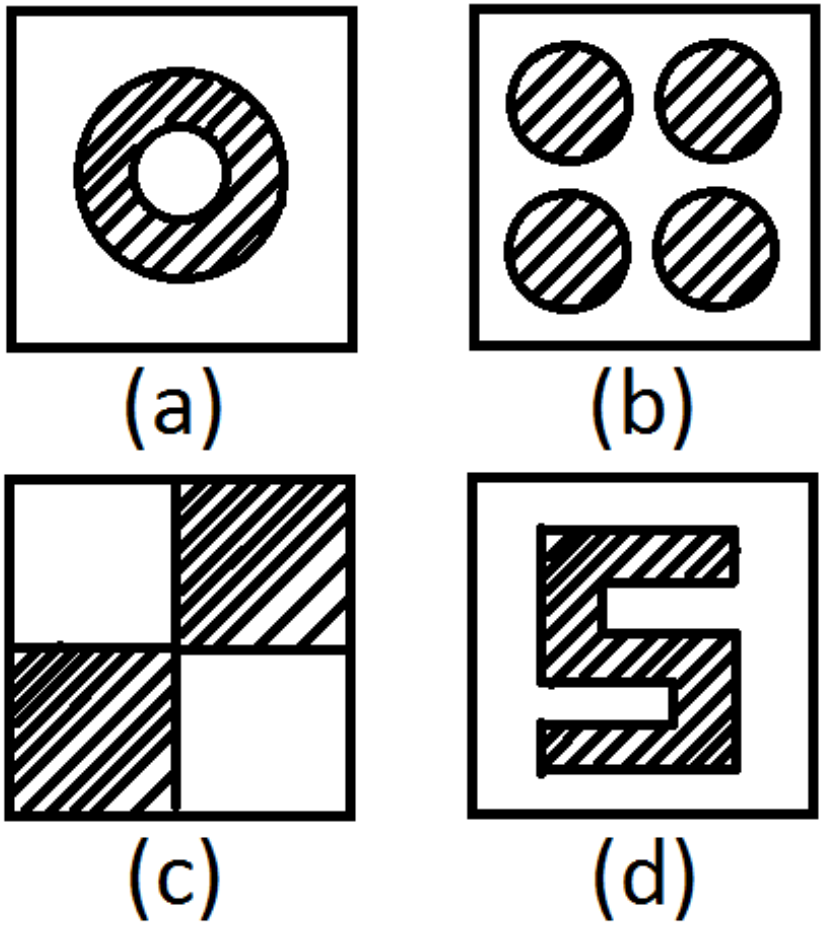


# On the Query Classification Dataset

| Classifier | Accuracy |
|---|---|
| Naive Bayes | $0.721 \pm 0.018$ |
| MaxEnt | $0.667 \pm 0.028$ |
| SVM | $0.735 \pm 0.032$ |
| Non-Linear 3-Layer Bayes | $0.711 \pm 0.032$ |

Table: Accuracy on query classification.

# Accuracy on Artificial Dataset



# Accuracy on Artificial Dataset

| Shape | GNB | GHB |
|---|---|---|
| ring | 0.664 | 0.949 |
| dots | 0.527 | 0.926 |
| XOR | 0.560 | 0.985 |
| S | 0.770 | 0.973 |

Table: Accuracy on the artificial dataset.

# Outline

# Conclusions

- We have shown that generative classifiers with a hidden layer are capable of learning non-linear decision boundaries under the right conditions (independence assumptions), and therefore might be said to be capable of deep learning.
- We have also shown experimentally that multinomial non-linear 3-layer Bayesian classifiers can outperform some linear classification algorithms on some text classification tasks.

# Errata

**In our survey of prior work ...**

Hierarchical Bayesian models are directed probabilistic graphical models in which the directed acyclic graph is a tree (each node that is not the root node has only one immediate parent). In these models the ancestors $a(i)$ of any node $i$ do not contain any siblings (every pair of ancestors is also in an ancestor-descendent relationship).

The above statement which is found in our paper is incorrect. Hierarchical Bayesian models including ours are generally speaking DAGs.

# Future Work

- Comparison of performance with deep neural networks; evaluation on image processing tasks; other training algorithms - Gibbs' sampling (MCMC), variational methods.
- Going from 3 layers to higher numbers of layers (solving the "vanishing information" problem).
- Integration with larger models, like sequential models (HMMs) and translation models (alignment models or sequence to sequence models).
- Probabilistic embeddings.
- Unified model covering Bayesian models and neural networks.

# Thank you!

Cohan Sujay Carlos & Geetanjali Rakshit

cohan@aiaioo.com