GENERATIVE MODELS AND EXPECTATION MAXIMIZATION

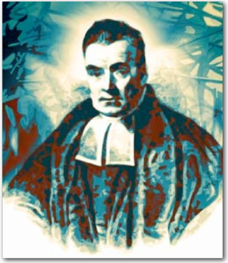
David Talbot, Yandex Translate

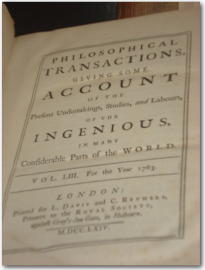
Autumn 2018

Yandex School of Data Analysis

CONDITIONAL PROBABILITY

REV. BAYES




PHILOSOPHICAL
TRANSACTIONS,
GIVING SOME
ACCOUNT
OF THE
Present Undertakings, Studies, and Labours,
OF THE
INGENIOUS,
IN MANY
Confiderable Parts of the WORLD.

VOL. LIII. For the Year 1763.

LONDON:
Printed for L. DAVIS and C. REYMERS,
Printers to the ROYAL SOCIETY,
against Gray's-Inn Gate, in Holbourn.
M.DCC.LXIV.

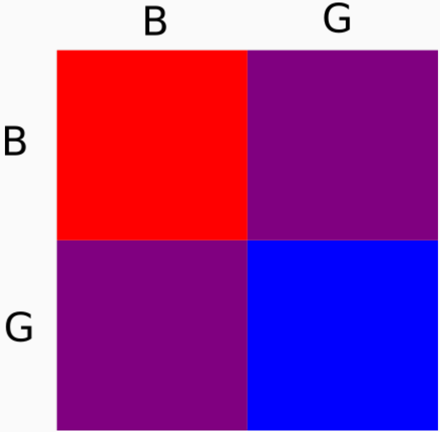
$$\Pr(X|Y) = \frac{\Pr(X)\Pr(Y|X)}{\Pr(Y)}$$

- Mr. White has two children. What is the probability that both children are boys?

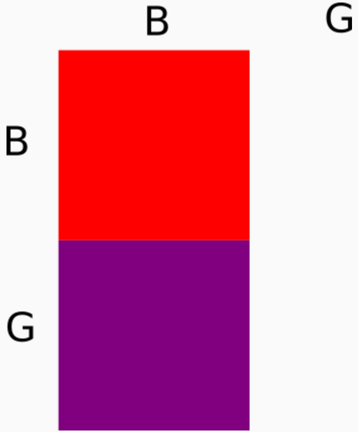
- Mr. White has two children. What is the probability that both children are boys?
- Mr. Jones has two children. The older child is a boy. What is the probability that both children are boys?

- Mr. White has two children. What is the probability that both children are boys?
- Mr. Jones has two children. The older child is a boy. What is the probability that both children are boys?
- Mr. Smith has two children. One of them is a boy. What is the probability that both children are boys?

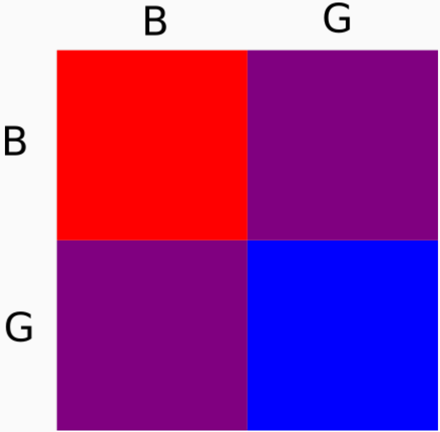
PRIOR PROBABILITY



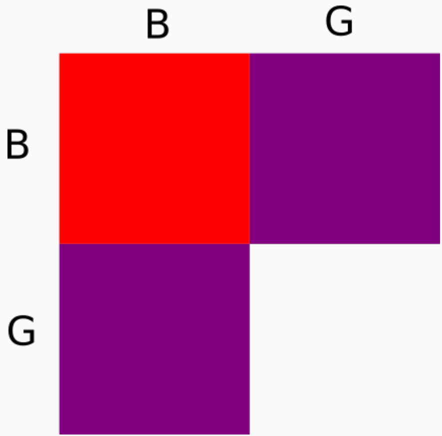
CONDITION ON EVENT ‘THE OLDER CHILD IS A BOY’



PRIOR PROBABILITY

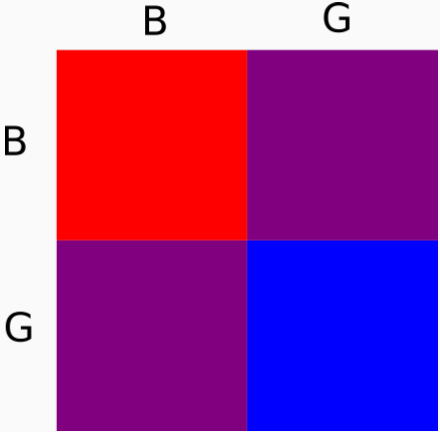


CONDITIONED ON THE EVENT 'ONE IS A BOY'

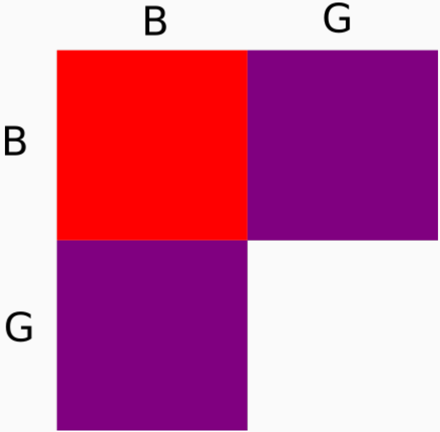


Mr. Brown has two children. One of them is a boy born on a Tuesday. What is the probability that he has two boys?

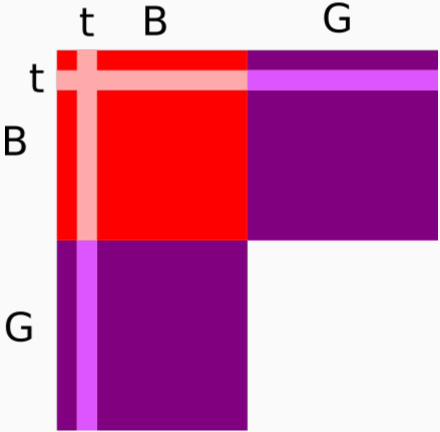
PRIOR PROBABILITY



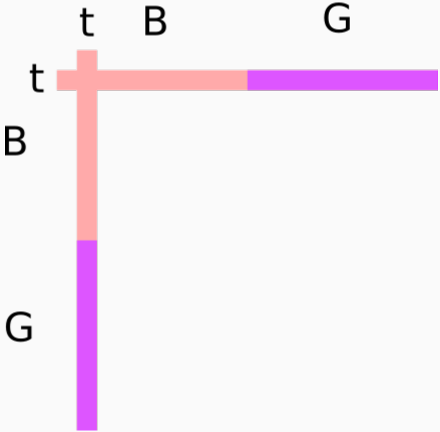
CONDITIONED ON 'ONE IS A BOY'



CONDITIONED ON ‘ONE IS A BOY BORN ON TUESDAY’



CONDITIONED ON 'ONE IS A BOY BORN ON TUESDAY'



HOW IS THE INFORMATION MADE AVAILABLE?

If ‘One is a boy’ is just an observation, then one model is:

HOW IS THE INFORMATION MADE AVAILABLE?

If ‘One is a boy’ is just an observation, then one model is:

- Draw a child $C_1 \in \{B, G\}$
- Draw a child $C_2 \in \{B, G\}$
- Draw an index $i \in \{1, 2\}$
- Observe that $i = 1$ and $C_1 = B$

HOW IS THE INFORMATION MADE AVAILABLE?

If ‘One is a boy’ is just an observation, then one model is:

- Draw a child $C_1 \in \{B, G\}$
- Draw a child $C_2 \in \{B, G\}$
- Draw an index $i \in \{1, 2\}$
- Observe that $i = 1$ and $C_1 = B$

$$\Pr(BB|C_1 = B) \quad =$$

If 'One is a boy' is just an observation, then one model is:

- Draw a child $C_1 \in \{B, G\}$
- Draw a child $C_2 \in \{B, G\}$
- Draw an index $i \in \{1, 2\}$
- Observe that $i = 1$ and $C_1 = B$

$$\Pr(BB|C_1 = B) = \frac{\Pr(BB)}{\Pr(C_1 = B)}$$

If ‘One is a boy’ is just an observation, then one model is:

- Draw a child $C_1 \in \{B, G\}$
- Draw a child $C_2 \in \{B, G\}$
- Draw an index $i \in \{1, 2\}$
- Observe that $i = 1$ and $C_1 = B$

$$\begin{aligned}
\Pr(BB|C_1 = B) &= \frac{\Pr(BB)}{\Pr(C_1 = B)} \\
&= \frac{\Pr(BB)}{\Pr(BB) + \Pr(BG)}
\end{aligned}$$

HOW IS THE INFORMATION MADE AVAILABLE?

If ‘One is a boy’ is just an observation, then one model is:

- Draw a child $C_1 \in \{B, G\}$
- Draw a child $C_2 \in \{B, G\}$
- Draw an index $i \in \{1, 2\}$
- Observe that $i = 1$ and $C_1 = B$

$$\begin{aligned} \Pr(BB|C_1 = B) &= \frac{\Pr(BB)}{\Pr(C_1 = B)} \\ &= \frac{\Pr(BB)}{\Pr(BB) + \Pr(BG)} \\ &= \frac{1}{2} \end{aligned}$$

HOW IS THE INFORMATION MADE AVAILABLE?

If ‘One is a boy’ is a prior constraint, then one model is

How is the information made available?

If ‘One is a boy’ is a prior constraint, then one model is

- While $B \notin \{C_1, C_2\}$ do:
  - Draw a child $C_1 \in \{B, G\}$
  - Draw a child $C_2 \in \{B, G\}$

HOW IS THE INFORMATION MADE AVAILABLE?

If 'One is a boy' is a prior constraint, then one model is

- While $B \notin \{C_1, C_2\}$ do:
  - Draw a child $C_1 \in \{B, G\}$
  - Draw a child $C_2 \in \{B, G\}$

$$\Pr(BB|B \in \{C_1, C_2\}) \quad =$$

HOW IS THE INFORMATION MADE AVAILABLE?

If ‘One is a boy’ is a prior constraint, then one model is

- While $B \notin \{C_1, C_2\}$ do:
  - Draw a child $C_1 \in \{B, G\}$
  - Draw a child $C_2 \in \{B, G\}$

$$\Pr(BB|B \in \{C_1, C_2\}) = \frac{\Pr(BB, B \in \{C_1, C_2\})}{\Pr(B \in \{C_1, C_2\})}$$

HOW IS THE INFORMATION MADE AVAILABLE?

If 'One is a boy' is a prior constraint, then one model is

- While $B \notin \{C_1, C_2\}$ do:
  - Draw a child $C_1 \in \{B, G\}$
  - Draw a child $C_2 \in \{B, G\}$

$$\begin{aligned}
\Pr(BB | B \in \{C_1, C_2\}) &= \frac{\Pr(BB, B \in \{C_1, C_2\})}{\Pr(B \in \{C_1, C_2\})} \\
&= \frac{\Pr(BB)}{\Pr(BB) + \Pr(BG) + \Pr(GB)}
\end{aligned}$$

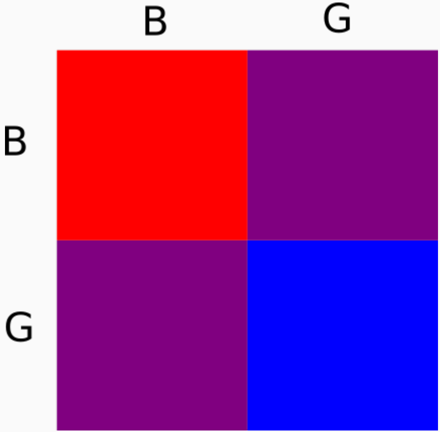
If 'One is a boy' is a prior constraint, then one model is

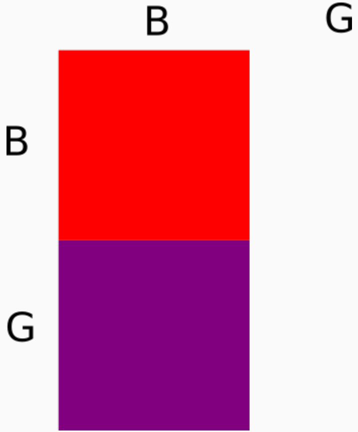
- While $B \notin \{C_1, C_2\}$ do:
  - Draw a child $C_1 \in \{B, G\}$
  - Draw a child $C_2 \in \{B, G\}$

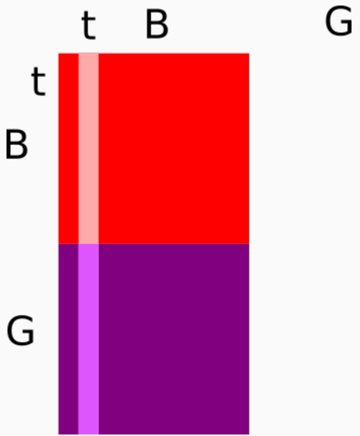
$$
\begin{aligned}
\Pr(BB|B \in \{C_1, C_2\}) &= \frac{\Pr(BB, B \in \{C_1, C_2\})}{\Pr(B \in \{C_1, C_2\})} \\
&= \frac{\Pr(BB)}{\Pr(BB) + \Pr(BG) + \Pr(GB)} \\
&= \frac{1}{3}
\end{aligned}
$$

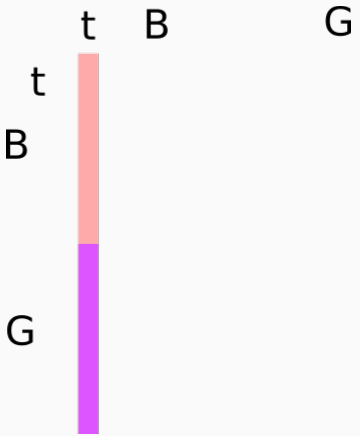
Mr. Brown has two children. One of them is a boy born on a Tuesday. What is the probability that he has two boys?

PRIOR PROBABILITY




B
G
B
G


t
B
G
t
B
G


t
B
G
t
B
G

Mr. Brown has two children. One of them is a boy born on Tuesday. What is the probability that both children are boys?

Mr. Brown has two children. One of them is a boy born on Tuesday. What is the probability that both children are boys?

- Draw a child $C_1 \in \{B_{mon}, B_{tue}, B_{wed}, \ldots, G_{mon}, G_{tue}, G_{wed}\}$
- Draw a child $C_2 \in \{B_{mon}, B_{tue}, B_{wed}, \ldots, G_{mon}, G_{tue}, G_{wed}\}$
- Draw an index $i \in \{1, 2\}$
- Observe that $i = 1$ and $C_1 = B_{tue}$

$$\Pr(BB|C_1 = B_{tue}) \quad = \quad \frac{\Pr(B_{tue}B)}{\Pr(B_{tue}B) + \Pr(B_{tue}G)}$$

$$\begin{aligned}
\Pr(BB|C_1 = B_{tue}) &= \frac{\Pr(B_{tue}B)}{\Pr(B_{tue}B) + \Pr(B_{tue}G)} \\
&= \frac{1/14 \times 1/2}{1/14 \times 1/2 + 1/14 \times 1/2}
\end{aligned}$$

$$
\begin{aligned}
\Pr(BB|C_1 = B_{tue}) &= \frac{\Pr(B_{tue}B)}{\Pr(B_{tue}B) + \Pr(B_{tue}G)} \\
&= \frac{1/14 \times 1/2}{1/14 \times 1/2 + 1/14 \times 1/2} \\
&= \frac{1}{2}
\end{aligned}
$$

If ‘One is a boy born on a Tuesday’ is a prior constraint, then:

- while $B_{tue} \notin \{C_1, C_2\}$ do
    - Draw a child $C_1 \in \{B_{mon}, B_{tue}, B_{wed}, \ldots, G_{mon}, G_{tue}, G_{wed}\}$
    - Draw a child $C_2 \in \{B_{mon}, B_{tue}, B_{wed}, \ldots, G_{mon}, G_{tue}, G_{wed}\}$

USING BAYES' RULE

$$\Pr(BB|B_t) \quad =$$

USING BAYES' RULE

$$
\begin{aligned}
\Pr(BB|B_t) &= \frac{\Pr(BB)\Pr(B_t|BB)}{\Pr(B_t)} \\
&=
\end{aligned}
$$

USING BAYES' RULE

$$
\begin{aligned}
\Pr(BB|B_t) &= \frac{\Pr(BB)\Pr(B_t|BB)}{\Pr(B_t)} \\
&= \frac{\Pr(BB)(1-\Pr(\neg t)^2)}{\Pr(BB)(1-\Pr(\neg t)^2)+\Pr(BG)\Pr(t)+\Pr(GB)\Pr(t)}
\end{aligned}
$$

$$\begin{aligned}
\Pr(BB|B_t) &= \frac{\Pr(BB)\Pr(B_t|BB)}{\Pr(B_t)} \\
&= \frac{\Pr(BB)(1-\Pr(\neg t)^2)}{\Pr(BB)(1-\Pr(\neg t)^2)+\Pr(BG)\Pr(t)+\Pr(GB)\Pr(t)} \\
&= \frac{1-(6/7)^2}{1-(6/7)^2+1/7+1/7}
\end{aligned}$$

USING BAYES' RULE

$$
\begin{aligned}
\Pr(BB|B_t) &= \frac{\Pr(BB)\Pr(B_t|BB)}{\Pr(B_t)} \\
&= \frac{\Pr(BB)(1-\Pr(\neg t)^2)}{\Pr(BB)(1-\Pr(\neg t)^2)+\Pr(BG)\Pr(t)+\Pr(GB)\Pr(t)} \\
&= \frac{1-(6/7)^2}{1-(6/7)^2+1/7+1/7} \\
&= \frac{13}{27}
\end{aligned}
$$

Let additional independent event X have probaility ϵ

$$\Pr(BB|B_X) \quad =$$

Let additional independent event X have probaility ϵ

$$\Pr(BB|B_X) = \frac{\Pr(BB)(1-(1-\epsilon)^2)}{\Pr(BB)(1-(1-\epsilon)^2)+\Pr(BG)\epsilon+\Pr(GB)\epsilon}$$

Let additional independent event X have probaility ϵ

$$
\begin{aligned}
\Pr(BB|B_X) &= \frac{\Pr(BB)(1-(1-\epsilon)^2)}{\Pr(BB)(1-(1-\epsilon)^2)+\Pr(BG)\epsilon+\Pr(GB)\epsilon} \\
&= \frac{1-(1-\epsilon)^2}{1-(1-\epsilon)^2+\epsilon+\epsilon}
\end{aligned}
$$

Let additional independent event X have probaility ϵ

$$
\begin{aligned}
\Pr(BB|B_X) &= \frac{\Pr(BB)(1-(1-\epsilon)^2)}{\Pr(BB)(1-(1-\epsilon)^2)+\Pr(BG)\epsilon+\Pr(GB)\epsilon} \\
&= \frac{1-(1-\epsilon)^2}{1-(1-\epsilon)^2+\epsilon+\epsilon} \\
&= \frac{2\epsilon-\epsilon^2}{4\epsilon-\epsilon^2}
\end{aligned}
$$

Let additional independent event X have probaility ϵ

$$
\begin{aligned}
\Pr(BB|B_X) &= \frac{\Pr(BB)(1-(1-\epsilon)^2)}{\Pr(BB)(1-(1-\epsilon)^2)+\Pr(BG)\epsilon+\Pr(GB)\epsilon} \\
&= \frac{1-(1-\epsilon)^2}{1-(1-\epsilon)^2+\epsilon+\epsilon} \\
&= \frac{2\epsilon-\epsilon^2}{4\epsilon-\epsilon^2} \\
&= \frac{2-\epsilon}{4-\epsilon}
\end{aligned}
$$

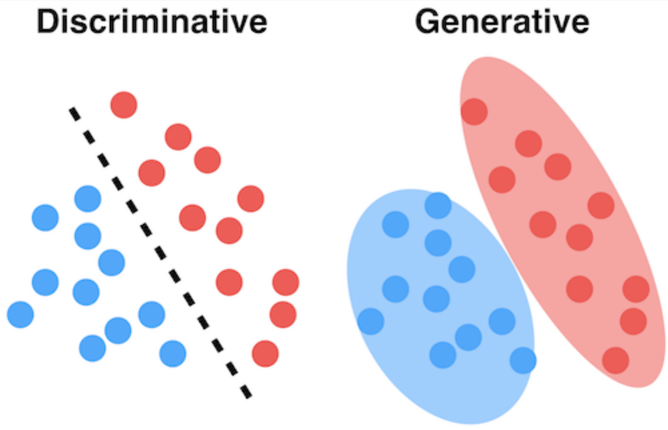
GENERATIVE MODELS

Generative models: a joint distribution over observations X and labels Y

$$\Pr(X, Y)$$

Discriminative models: a conditional distribution over the labels

$$\Pr(Y|X)$$


Discriminative
Generative

Your friend has a bag of different coloured coins.

- She draws a coin at random
- She tosses the coin n times

$$X \in \{H, T\}^n$$

$$Y \in \{R, O, Y, G, B, I, V\}$$

Assuming that coins of the same colour are identical

- What *parameters* describe a *generative model* of this data?
- What *statistics* do we need to estimate these parameters?
- What are the *maximum likelihood estimates* for these parameters?

Choose parameters $\lambda, \theta_R, \theta_b$ s.t. *likelihood* of the data X is maximized, i.e.

$$\theta^* = \operatorname*{argmax}_\theta \Pr(X|\theta).$$

Often easier to work with logarithm, e.g.

$$\log \Pr(R, H, H, T) = \log P(R) + \log \Pr(H, H, T|R).$$

So we can find the maximum of each parameter separately.

We observed a sample D drawn from $(x, y) \in (X, Y)$ where $X \in \{H, T\}$, $Y = \{R, B\}$. Each observation was labeled so,

$$\begin{aligned}\hat{\theta}_{mle} &= \underset{\theta}{\operatorname{argmax}} \sum_{(x,y) \in D} \log \Pr(X = x, Y = y | \theta) \\ &= \underset{\theta}{\operatorname{argmax}} \sum_{(x,y) \in (X,Y)} \#(X = x, Y = y) \log \Pr(X = x, Y = y | \theta)\end{aligned}$$

where we summarized the data using the *sufficient statistics*.

MAXIMUM LIKELIHOOD ESTIMATES FOR OUR MODEL

$$
\begin{aligned}
\Pr(R) \qquad \lambda &= \frac{\#(R)}{\#(R) + \#(B)} \\
\Pr(H|R) \qquad \theta_R &= \frac{\#(H,R)}{\#(R)} \\
\Pr(H|B) \qquad \theta_B &= \frac{\#(H,B)}{\#(B)}
\end{aligned}
$$

If $T(X)$ are *sufficient statistics* for the sample X with respect to a model with parameters θ then

$$\Pr(\theta|T(X)) = \Pr(\theta|X).$$

Sufficient statistics summarize all the information about a sample that can influence our estimate of the parameters.

INDEPENDENCE ASSUMPTIONS

Generative models often make *independence assumptions*.

Generative models often make *independence assumptions*.

Why?

Generative models often make *independence assumptions*.

Why?

- Naive Bayes spam filter

Generative models often make *independence assumptions*.

Why?

- Naive Bayes spam filter: each word is generated independently given the class $\{S, \neg S\}$.

Generative models often make *independence assumptions*.

Why?

- Naive Bayes spam filter: each word is generated independently given the class $\{S, \neg S\}$.
- N-gram language model

Generative models often make *independence assumptions*.

Why?

- Naive Bayes spam filter: each word is generated independently given the class $\{S, \neg S\}$.
- N-gram language model: each word is generated independently given the $N - 1$ preceding words.

Generative models often make *independence assumptions*.

Why?

- Naive Bayes spam filter: each word is generated independently given the class $\{S, \neg S\}$.
- N-gram language model: each word is generated independently given the $N - 1$ preceding words.
- HMM POS tagger:

Generative models often make *independence assumptions*.

Why?

- Naive Bayes spam filter: each word is generated independently given the class $\{S, \neg S\}$.
- N-gram language model: each word is generated independently given the $N - 1$ preceding words.
- HMM POS tagger:each word is generated independently given its tag.

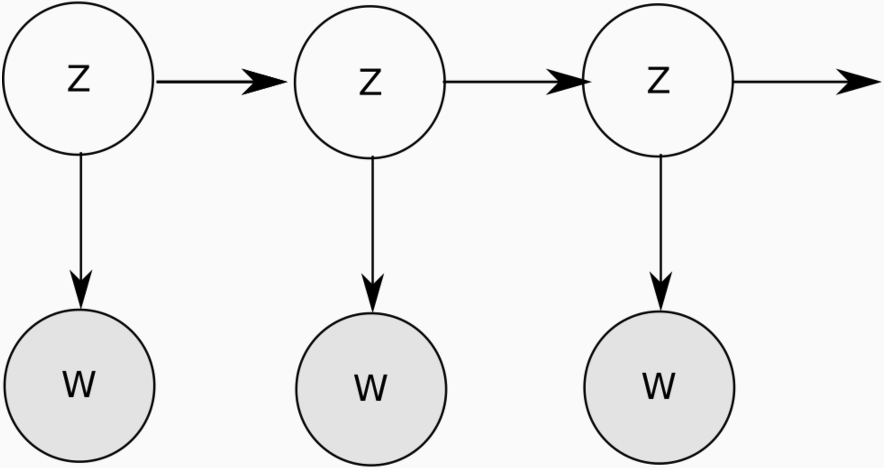
Generative models often make *independence assumptions*.

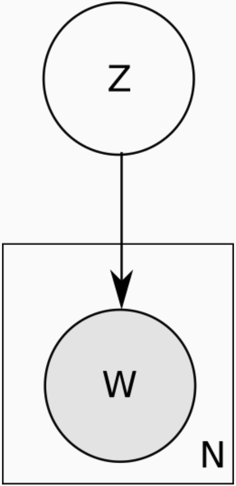
Why?

- Naive Bayes spam filter: each word is generated independently given the class $\{S, \neg S\}$.
- N-gram language model: each word is generated independently given the $N - 1$ preceding words.
- HMM POS tagger:each word is generated independently given its tag.

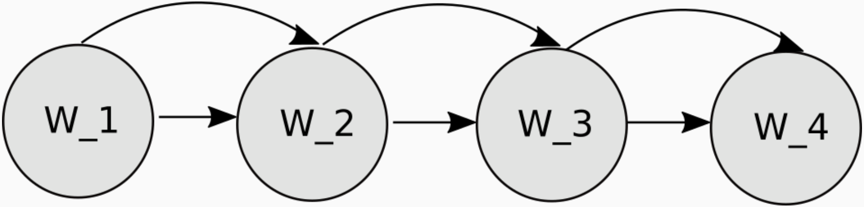
What are the sufficient statistics for each of these models?

GENERATIVE MODELS: HIDDEN MARKOV MODEL




Z
W
N

GENERATIVE MODELS: BIGRAM MODEL



Your careless friend dropped the bag of coins in the bath.

The paint wasn't waterproof so the coins are now identical...

How would you estimate the parameters now?

i.e. you see only (H, H, H), (T, T, H), (H, T, T), (H, H, T), (H, T, T).

EXPECTATION-MAXIMIZATION

EM MAXIMIZES A BOUND ON THE OBSERVED LIKELIHOOD

$$\begin{aligned}
\log \Pr(X|\theta) &= \log \sum_Z \Pr(X, Z|\theta) \\
&= \log \sum_Z q(Z) \frac{\Pr(X, Z|\theta)}{q(Z)} \\
&\geq \sum_Z q(Z) \log \frac{\Pr(X, Z|\theta)}{q(Z)} \\
&= \sum_Z q(Z) \log \Pr(X, Z|\theta) - \sum_Z q(Z) \log q(Z) \\
&= \sum_Z q(Z) \log \Pr(X, Z|\theta) + H(Z)
\end{aligned}$$

If $q(Z)$ does not depend on θ we can ignore the $H(x)$ term.

$$
\begin{aligned}
\log \Pr(X|\theta) &\geq \sum_{Z} q(Z) \log \frac{\Pr(X, Z|\theta)}{q(Z)} \\
&\geq \sum_{Z} q(Z) \log \frac{\Pr(X|\theta)\Pr(Z|X, \theta)}{q(Z)} \\
&= \sum_{Z} q(Z) \log \Pr(X|\theta) - \sum_{Z} q(Z) \log \frac{q(Z)}{\Pr(Z|, X, \theta)} \\
&= \log \Pr(X|\theta) - KL(q(Z)||\Pr(Z|, X, \theta))
\end{aligned}
$$

which implies that if $q(Z) = \Pr(Z|, X, \theta)$ the bound is tight.

We observed a sample D drawn from $(x, z) \in (X, Z)$ where $X \in \{H, T\}$, $Z = \{Red, Blue\}$. Each observation was labeled so,

$$\begin{aligned} \hat{\theta}_{mle} &= \underset{\theta}{\operatorname{argmax}} \sum_{(x,z)\in D} \log \Pr(X = x, Z = z|\theta) \\ &= \underset{\theta}{\operatorname{argmax}} \sum_{(x,z)\in(X,Z)} \#(X = x, Z = z) \log \Pr(X = x, Z = z|\theta) \end{aligned}$$

where we summarized the data using the *sufficient statistics*.

Let's reformulate the expression for *mle* estimation.

$$\hat{\theta}_{mle} = \underset{\theta}{\operatorname{argmax}} \sum_{(x,z)\in(X,Z)} \#(X = x, Z = z) \log \Pr(X = x, Z = z|\theta)$$
$$= \underset{\theta}{\operatorname{argmax}} \sum_{(x,z)\in D} \sum_{y\in\{Red,Blue\}} \delta(z, y) \log \Pr(X = x, Z = z|\theta)$$

where $\delta(x, y) = 1 \iff x = y$ otherwise 0.

We observed a sample D drawn from $(x, z) \in (X, Z)$ where $X \in \{H, T\}$, $Z = \{Red, Blue\}$. This time Z is hidden.

$$\hat{\theta}_{mle} = \underset{\theta}{\operatorname{argmax}} \sum_{(x,z) \in D} \sum_{y \in \{Red, Blue\}} \delta(z, y) \log \Pr(X = x, Z = z | \theta)$$

We observed a sample D drawn from $(x, z) \in (X, Z)$ where $X \in \{H, T\}$, $Z = \{Red, Blue\}$. This time Z is hidden.

$$\hat{\theta}_{mle} = \operatorname*{argmax}_{\theta} \sum_{(x,z) \in D} \sum_{y \in \{Red, Blue\}} \delta(z, y) \log \Pr(X = x, Z = z | \theta)$$

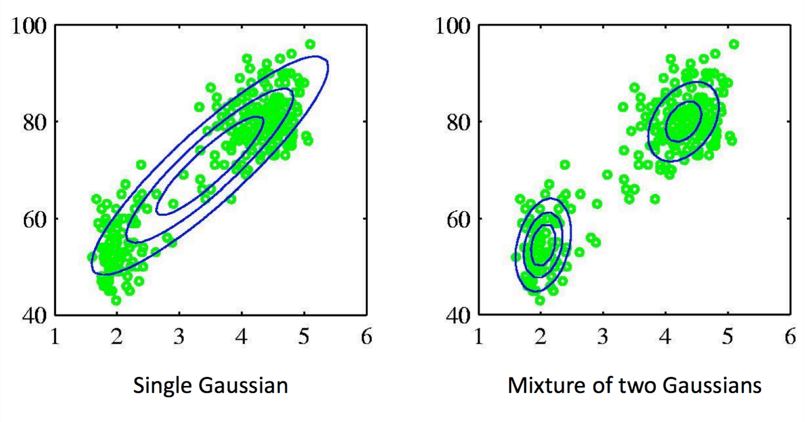
Replace $\delta(z, y) \in \{0, 1\}$ by our best guess $\Pr(Z = z | X = x, \theta_i)$.

$$\hat{\theta}_{i+1} = \operatorname*{argmax}_{\theta} \sum_{x \in D} \sum_{z \in \{Red, Blue\}} \Pr(Z = z | X = x, \theta_i) \log \Pr(X = x, Z = z | \theta_i)$$

This term is known as the *expected log-likelihood*.

- Initialize the parameters θ_0 somehow (randomly?)
- E-step: Compute $\Pr(Z|X, \theta_i)$ i.e. our best guess of the hidden data Z given our current parameters. (Think of $\Pr(Z|X, \theta_i)$ as a fractional count of Z.)
- M-step: Update the parameters θ_{i+1} to maximize the expected log-likelihood.
- Iterate until the expected log-likelihood stops increasing.

Intuition: if we knew θ we could just infer Z (usually), likewise if we knew Z we could just estimate θ (you did this). Since we don't know either, just guess and iteratively improve.

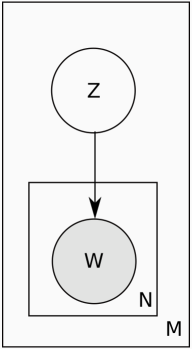

100
80
60
40
1
2
3
4
5
6
Single Gaussian
100
80
60
40
1
2
3
4
5
6
Mixture of two Gaussians

1. Choose a cluster $i \in \{1, 2, \ldots, K\}$ from prior $\Pr(Y = i) = \lambda_i$
2. Generate an observation X from a Gaussian g_i with parameters μ_i, σ_i

$$\Pr(X = x|\theta) = \sum_{i \in \{1,2,\ldots,K\}} \Pr(Y = i)\Pr(X = x|Y = i) = \sum_{i \in \{1,2,\ldots,K\}} \lambda_i g_i(x)$$

How does a mixture model improve on a single Gaussian model?

TOPIC MIXTURE MODEL

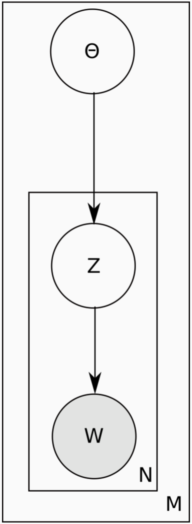


- Sample a topic $z \in \{1, 2, \ldots, K\}$ for a document
- Generate words independently given the topic

$$\Pr(W_1, W_2, \ldots, W_N) = \prod_{i=1}^{N} \Pr(W_i = w | Z = z)$$

How can the topic variable help here?

LDA OR ADMIXTURE MODEL



- Sample a distribution over topics for a document

$$\theta = (\theta_1, \theta_2, \ldots, \theta_K) \sim \text{Dirichlet}(\alpha)$$

For each word in the document:

- Generate a topic Z for a word

$$Z_i = \Pr(Z_i = z) = \theta_z$$

- Generate a word W according to the topic distribution

$$W_i = \Pr(W_i = w | Z = z) = \beta_{z,w}$$

1. Estimate the posterior probability of each cluster for each data point (E-step)
2. Update parameters using these posterior probabilities as fractional counts (M-step)

$$\Pr(Y = i | X = x, \theta) = \frac{p_i g_i(x)}{\sum_{j \in \{1,2,\ldots,k\}} p_j g_j(x)}$$

[K-means]

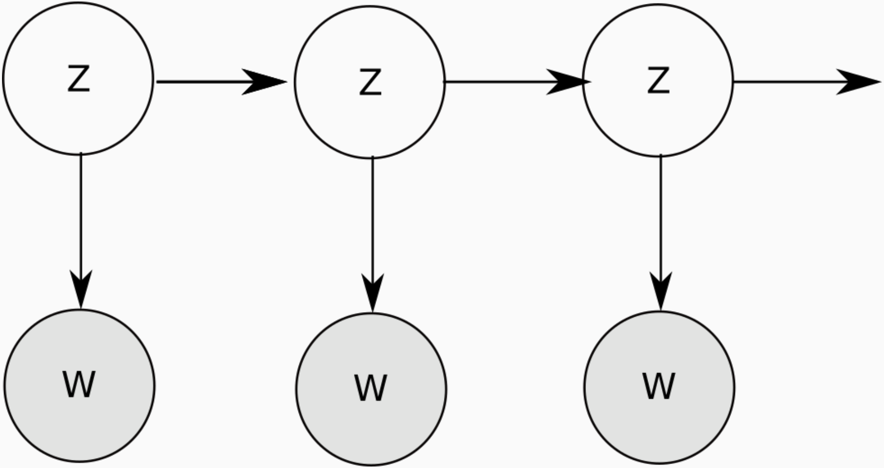
Assign data to cluster with highest posterior (e.g. hard EM)

$$i^* = \underset{i}{\operatorname{argmax}} \Pr(Cluster = i | Data = x)$$

[Gibbs]

Sample a cluster assignment from the posterior

$$i^* \sim \Pr(Cluster = i | Data = x)$$



Useful for tagging, segmentation, speech, etc.

HIDDEN MARKOV MODEL: PARAMETERS

Parameters:

$$\theta = (\pi, A, O)$$

Parameters:

$$\theta = (\pi, A, O)$$

Probability of starting in state i:

$$\pi_i = \Pr(Z_0 = i)$$

Parameters:

$$\theta = (\pi, A, O)$$

Probability of starting in state i:

$$\pi_i = \Pr(Z_0 = i)$$

Probability of moving from state i to j:

$$A_i(j) = \Pr(Z_t = j | Z_{t-1} = i)$$

Parameters:

$$\theta = (\pi, A, O)$$

Probability of starting in state i:

$$\pi_i = \Pr(Z_0 = i)$$

Probability of moving from state i to j:

$$A_i(j) = \Pr(Z_t = j | Z_{t-1} = i)$$

Probability of emitting x given we're in state i:

$$O_i(x) = \Pr(X_t = x | Z_t = i)$$

Parameters:

$$\theta = (\pi, A, O)$$

Probability of starting in state i:

$$\pi_i = \Pr(Z_0 = i)$$

Probability of moving from state i to j:

$$A_i(j) = \Pr(Z_t = j | Z_{t-1} = i)$$

Probability of emitting x given we're in state i:

$$O_i(x) = \Pr(X_t = x | Z_t = i)$$

What are the independence assumptions?

Parameters:

$$\theta = (\pi, A, O)$$

Probability of starting in state i:

$$\pi_i = \Pr(Z_0 = i)$$

Probability of moving from state i to j:

$$A_i(j) = \Pr(Z_t = j | Z_{t-1} = i)$$

Probability of emitting x given we're in state i:

$$O_i(x) = \Pr(X_t = x | Z_t = i)$$

What are the independence assumptions?

What are the sufficient statistics?

In the observed case, we need the following statistics:

$$\#(Z_0 = i)$$

$$\#(Z_{t-1} = i, Z_t = j)$$

$$\#(X_t = x, Z_t = i)$$

In the hidden case, we need expectations for each sample:

$$\#(Z_0 = i) \to \Pr(Z_0 = i | X_{0:T} = x_{0:T}, \theta)$$

$$\#(Z_{t-1} = i, Z_t = j) \to \Pr(Z_{t-1} = i, Z_t = j | X_{0:T} = x_{0:T}, \theta)$$

$$\#(X_t = x, Z_t = i) \to \Pr(Z_t = i | X_{0:T} = x_{0:T}, \theta) \#(X_t = x)$$

We want to compute:

$$\Pr(Z_t = z | X_{0:T} = x_{0:T}, \theta) = \frac{\Pr(Z_t = z, X_{0:T} = x_{0:T})}{\Pr(X_{0:T} = x_{0:T})}$$

HIDDEN MARKOV MODEL: COMPUTING POSTERIOR

We want to compute:

$$\Pr(Z_t = z | X_{0:T} = x_{0:T}, \theta) = \frac{\Pr(Z_t = z, X_{0:T} = x_{0:T})}{\Pr(X_{0:T} = x_{0:T})}$$

But the computation looks exponential in the length T ...

$$\Pr(X_{0:T} = x_{0:T}) = \sum_{z_0} \sum_{z_1} \cdots \sum_{z_{T-1}} \sum_{z_T} \Pr(x_{0:T}, z_0, z_1, \ldots, z_T | \theta)$$

Use HMM independence assumptions to factorize

$$\Pr(x_0, \ldots, x_t, z_t, x_{t+1}, \ldots, x_T|\theta) = \Pr(x_0, \ldots, x_t, z_t|\theta)\Pr(x_{t+1}, \ldots, x_T|z_t, \theta).$$

Use HMM independence assumptions to factorize

$\Pr(x_0, \ldots, x_t, z_t, x_{t+1}, \ldots, x_T|\theta) = \Pr(x_0, \ldots, x_t, z_t|\theta)\Pr(x_{t+1}, \ldots, x_T|z_t, \theta).$

If we can compute this, then the denominator is easy

$$\Pr(x_0, \ldots, x_T|\theta) = \sum_{z_t} \Pr(x_0, \ldots, x_t, z_t|\theta)\Pr(x_{t+1}, \ldots, x_T|z_t, \theta).$$

Compute $\Pr(x_0, \ldots, x_t, z_t|\theta)$ from $\Pr(x_0, \ldots, x_{t-1}, z_{t-1}|\theta)$ as, Compute $\Pr(x_0, \ldots, x_t, z_t|\theta)$ from $\Pr(x_0, \ldots, x_{t-1}, z_{t-1}|\theta)$ as,

$$\begin{aligned}
\Pr(x_0, \ldots, x_t, z_t|\theta) &= \sum_{z_{t-1}} \Pr(x_0, \ldots, x_{t-1}, x_t, z_{t-1}, z_t|\theta) \\
&= \sum_{z_{t-1}} \Pr(x_0, \ldots, x_{t-1}, z_{t-1}|\theta)\Pr(z_t|z_{t-1})\Pr(x_t|z_t) \\
&= \sum_{z_{t-1}} \Pr(x_0, \ldots, x_{t-1}, z_{t-1}|\theta)A_{z_{t-1}}(z_t)O_{z_t}(x_t)
\end{aligned}$$

Definition:

$$\alpha_t(z) \equiv \Pr(x_0, \ldots, x_t, z_t | \theta)$$

Initialization:

$$\alpha_0(i) = \pi_i O_i(x_0)$$

Recursion:

$$\alpha_{t+1}(i) = \sum_j \alpha_t(j) A_j(i) O_i(x_t)$$

Gives us the probability of observed sequence since,

$$\Pr(x_0, \ldots, x_T | \theta) = \sum_{z_T} \Pr(x_0, \ldots, x_T, z_T | \theta) = \sum_i \alpha_T(i).$$

Definition:

$$\beta_t(z) \equiv \Pr(x_{t+1}, \ldots, x_T | z_t, \theta)$$

Initialization:

$$\beta_T(i) = 1$$

Recursion:

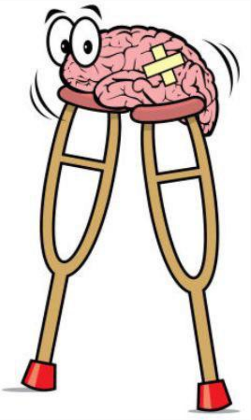
$$\beta_t(i) = \sum_j \beta_{t+1}(j) A_i(j) O_i(x_{t+1})$$

Posterior probabilities over single states

$$
\begin{aligned}
\Pr(Z_t = i | x_0, \ldots, x_T; \theta) &= \frac{\Pr(Z_t = i, x_0, \ldots, x_T | \theta)}{\Pr(x_0, \ldots, x_T | \theta)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_j \alpha_t(j)\beta_t(j)}
\end{aligned}
$$

Posterior probabilities over state transitions

$$\begin{aligned}
\Pr(Z_t = i, Z_{t+1} = j | x_0, \ldots, x_T; \theta) = & \quad \frac{\Pr(Z_t = i, Z_{t+1} = j, x_0, \ldots, x_T | \theta)}{\Pr(x_0, \ldots, x_T | \theta)} \\
= & \quad \frac{\alpha_t(i) A_j(i) O_i(x_{t+1}) \beta_{t+1}(i)}{\sum_i \sum_j \alpha_t(i) A_j(i) O_i(x_{t+1}) \beta_{t+1}(i)}
\end{aligned}$$



Initialize complex models with parameters from simpler ones.