GENERATIVE MODELS AND HIDDEN VARIABLES

David Talbot, Yandex Translate

Autumn 2019

Yandex School of Data Analysis

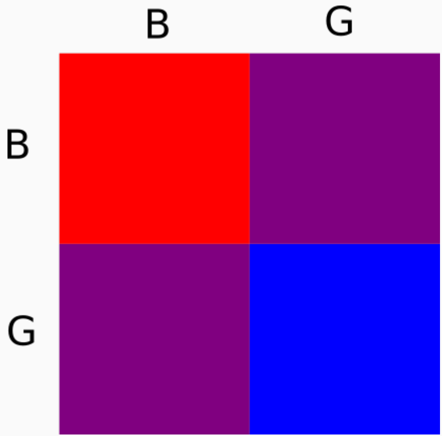
PRIOR AND CONDITIONAL PROBABILITY

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

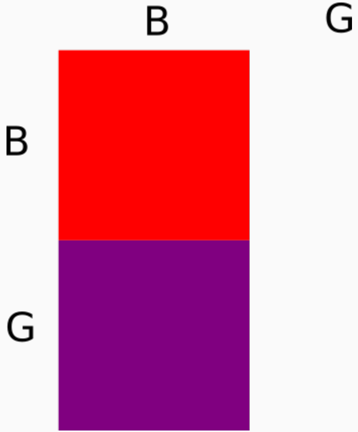
- Mr. White has two children. What is the probability that both children are boys?
- Mr. Jones has two children. The older child is a boy. What is the probability that both children are boys?

- 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?

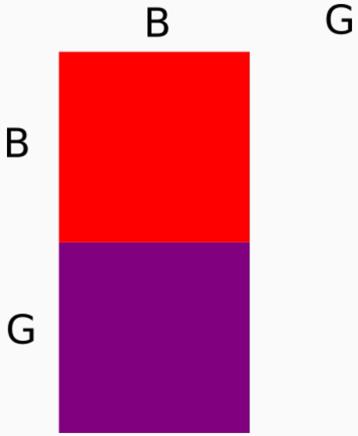
PRIOR PROBABILITY



CONDITION ON EVENT ‘THE OLDER CHILD IS A BOY’

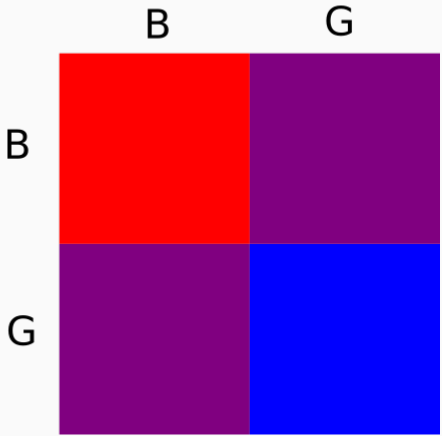


CONDITION ON EVENT 'THE OLDER CHILD IS A BOY'

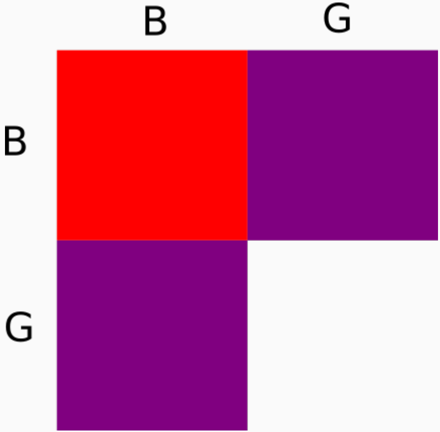


So $\Pr(BB) = \frac{1}{2}$

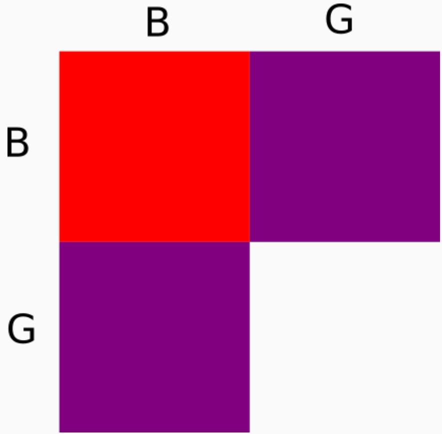
PRIOR PROBABILITY



CONDITIONED ON THE EVENT ‘ONE IS A BOY’



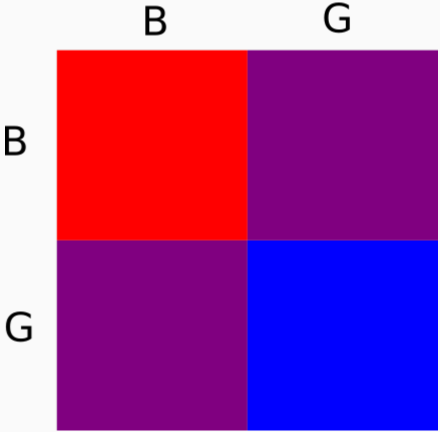
CONDITIONED ON THE EVENT 'ONE IS A BOY'



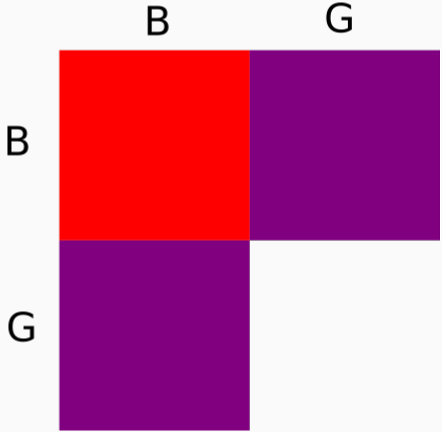
So $\Pr(BB) = \frac{1}{3}$

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

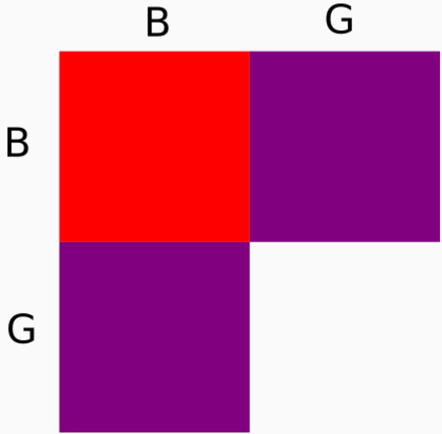
PRIOR PROBABILITY



CONDITIONED ON 'ONE IS A BOY'

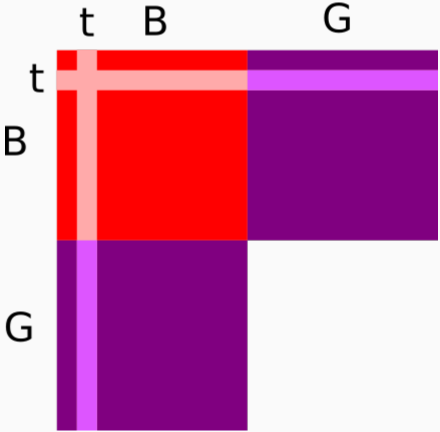


CONDITIONED ON 'ONE IS A BOY'

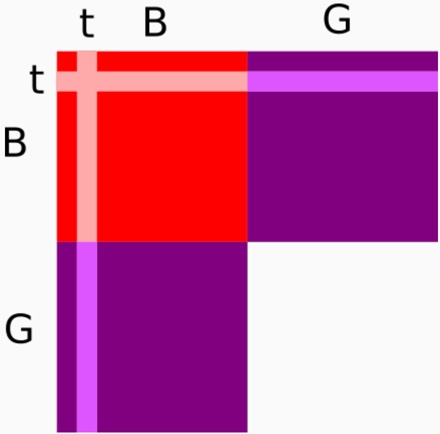


So $\Pr(BB) = \frac{1}{3}$

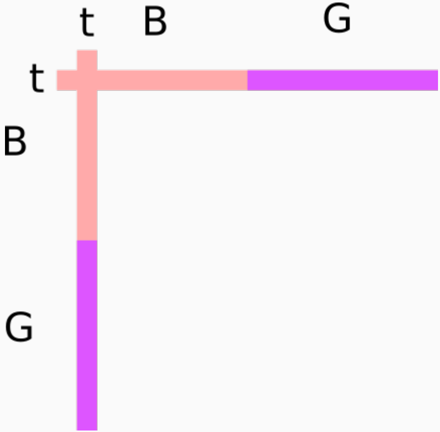
CONDITIONED ON 'ONE IS A BOY BORN ON TUESDAY'



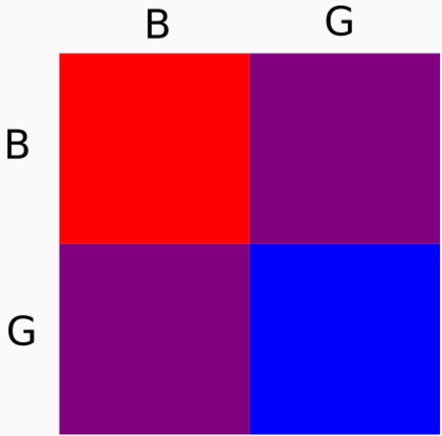
CONDITIONED ON 'ONE IS A BOY BORN ON TUESDAY'



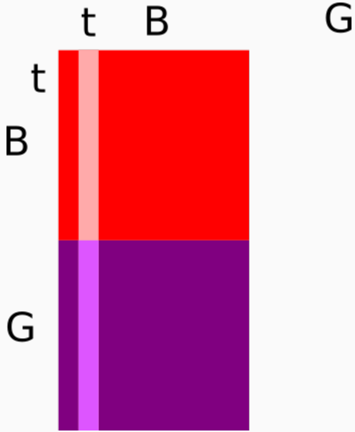
So $\Pr(BB) = \frac{13}{27} \approx \frac{1}{2}$


t
B
G
t
B
G

PRIOR PROBABILITY



CONDITIONAL PROBABILITY


t
B
G
t
B
G


t
B
G
t
B
G

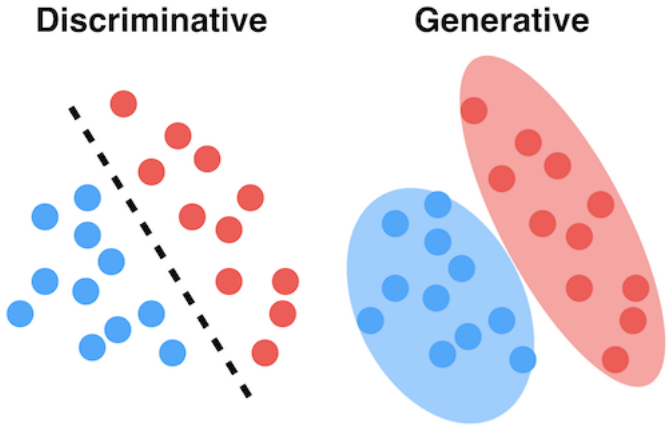
GENERATIVE MODELS

Generative models: joint distribution over X and Y

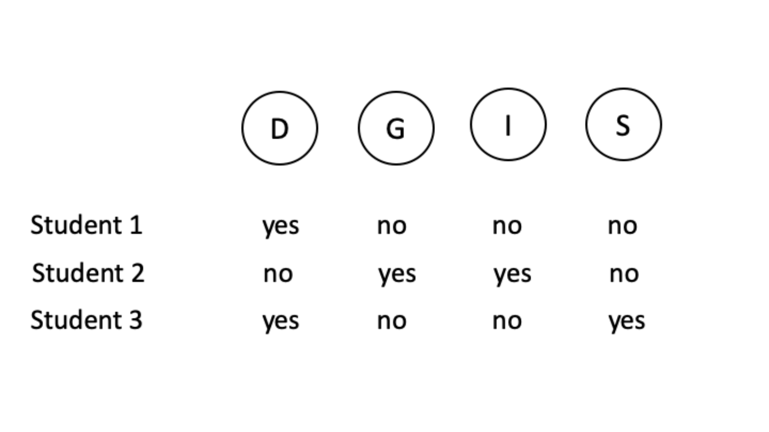
$$\Pr(X, Y|\theta).$$

Discriminative models: conditional distribution over Y

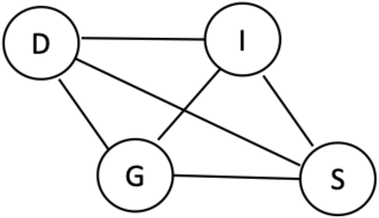
$$\Pr(Y|X, \theta).$$


Discriminative
Generative

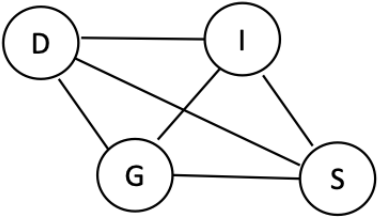
Students' grades: D = difficulty, I = intelligence, G = grade, S = SAT score



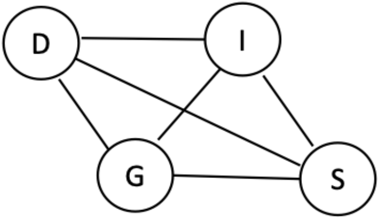
D = difficulty, I = intelligence, G = grade, S = SAT score



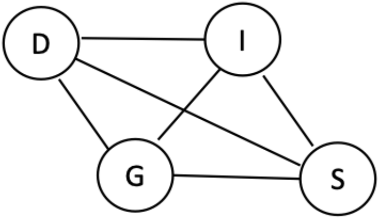
How many parameters needed if variables are *categorical*?



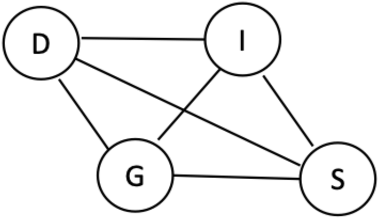
$|D| \times |I| \times |G| \times |S|$



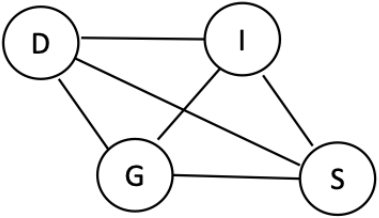
Assuming binary variables, parameters are $\Pr(D = yes, I = yes, G = no, S = yes)$ etc.



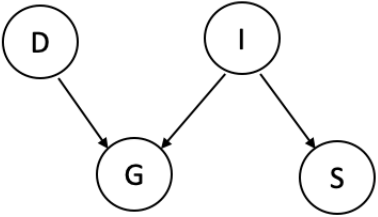
X^V parameters if V variables each with X values



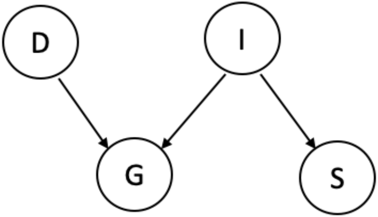
How can we reduce the number of parameters?



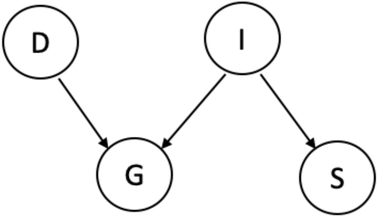
What independence assumptions does this model make?



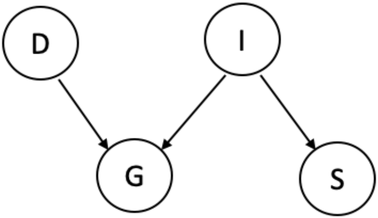
$$\Pr(D, I, G, S) = \Pr(D)\Pr(I)\Pr(G|D, I)\Pr(S|I)$$



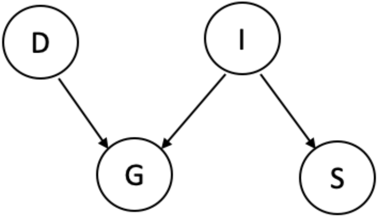
How many parameters are left?



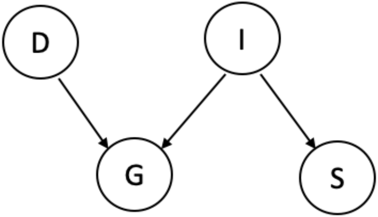
How could we reduce this further?



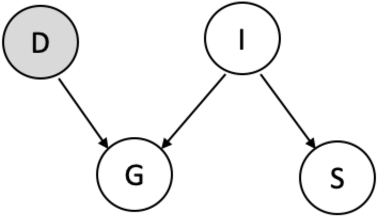
Is G independent of S in this model *a priori*?



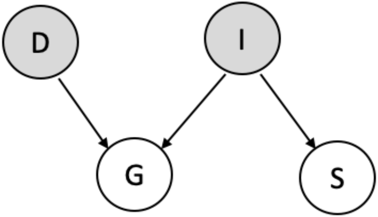
Compute $\Pr(S)$ and $\Pr(G)$. What can you conclude?



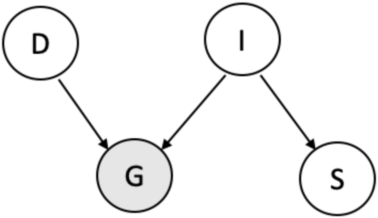
How does independence between variables change?



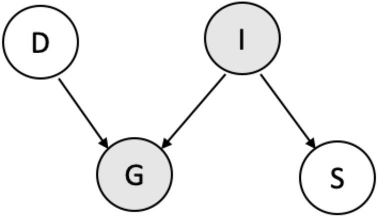
How does independence between variables change?



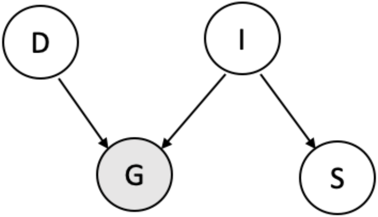
How does independence between variables change?



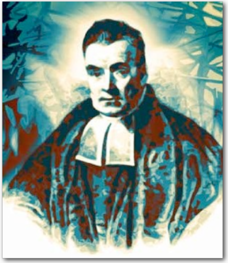
How does independence between variables change?

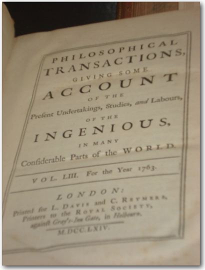


How difficult was the exam given only G?



REV. BAYES





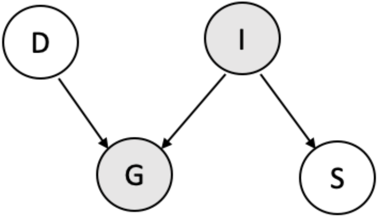
PHILOSOPHICAL
TRANSACTIONS,
GIVING SOME
ACCOUNT
OF THE
Prefent Undertakings, Studies, *and* Labours,
OF THE
INGENIOUS,
IN MANY
Confiderable Parts of the WORLD.

VOL. LIII. For the Year 1763.

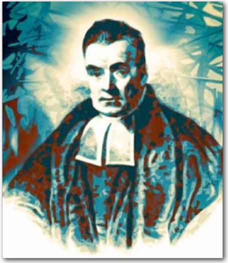
LONDON:
Printed for L. DAVIS and C. REYMERS,
Printers to the ROYAL SOCIETY,
against *Gray's-Inn Gate*, in *Holbourn*.
M.DCC.LXIV.

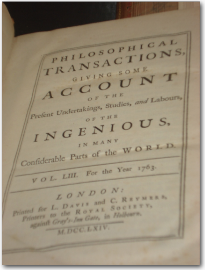
$$\Pr(X|Y) = \frac{\Pr(X)\Pr(Y|X)}{\Pr(Y)}$$

How difficult was the exam given G and I?



REV. BAYES





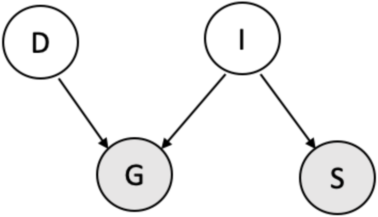
PHILOSOPHICAL
TRANSACTIONS,
GIVING SOME
ACCOUNT
OF THE
Present Undertakings, Studies, and Labours,
OF THE
INGENIOUS,
IN MANY
Confiderable Parts of the WORLD.

VOL. LIII. For the Year 1763.

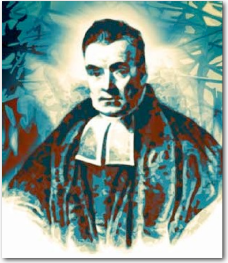
LONDON:
Printed for L. Davis and C. Reymers,
Printers to the Royal Society,
against Gray's-Inn Gate, in Holbourn.
M.DCC.LXIV.

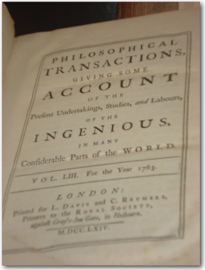
$$\Pr(X|Y) = \frac{\Pr(X)\Pr(Y|X)}{\Pr(Y)}$$

How difficult was the exam given G and S?



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.

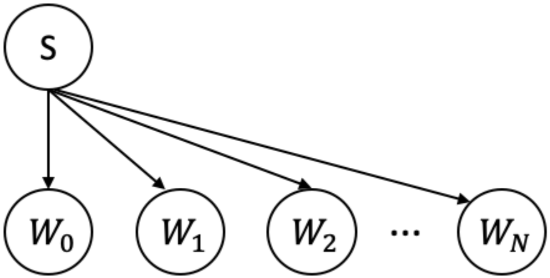
$$\Pr(X|Y) = \frac{\Pr(X)\Pr(Y|X)}{\Pr(Y)}$$

Some document models

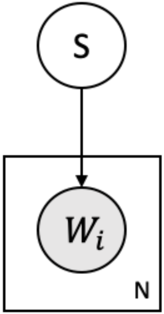
- Bag of Words model (aka Naive Bayes)
- Bigram Topic Model
- Latent Dirichlet Allocation (later)
- Hidden Markov model (later)

Strong independence assumptions

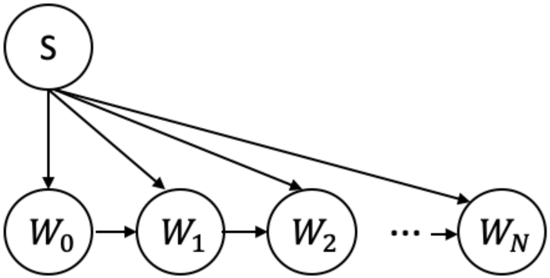
How can this possibly work for Spam detection?



How can this possibly work for Spam detection?



Why is this not be a good model for Spam detection?



Your friend has a bag of coins of different colours.

- They draw a coin at random
- They toss 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.

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

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*.

MAXIMIMUM 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.

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).

HIDDEN VARIABLE MODELS

- Hidden variables may or may not have a physical meaning

- Attributes may be unobserved on some examples (e.g. due to problems with data collection)

- Attributes may be hard to measure (e.g. intelligence)

- Sometimes adding a hidden variable simplifies a model ...

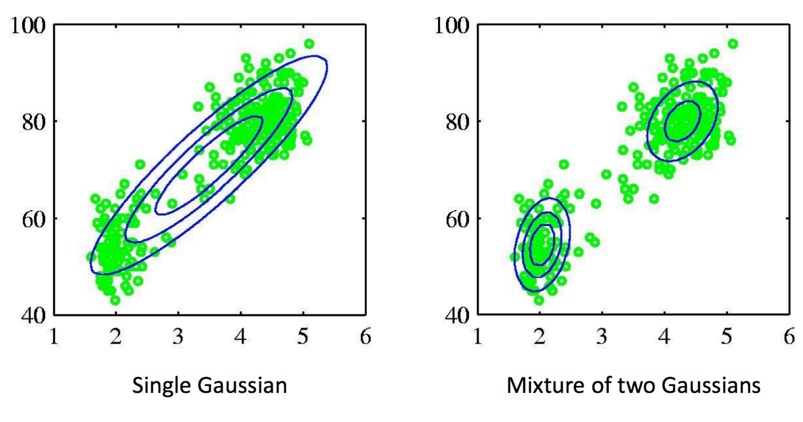
- Bigram language model (no hidden variables)

$$\Pr(w_t | w_{t-1}, \dots, w_0) \approx \Pr(w_t | w_{t-1})$$

- Class-based language model

$$\Pr(w_t | w_{t-1}, \dots, w_0) \approx \Pr(w_t | C(w_{t-1}))$$

MIXTURE MODELS

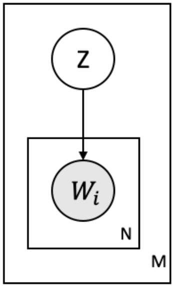


1. Choose a cluster $i \in \{1, 2, \ldots, K\}$ from prior $\Pr(Y = i) = \lambda_i$
2. Generate an observation X from a Gaussian g_i with parameters μ_i, σ_i

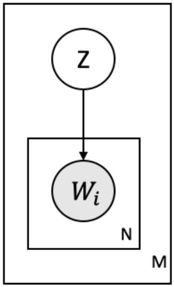
$$\Pr(X = x|\theta) = \sum_{i \in \{1,2,\ldots,K\}} \Pr(Y = i)\Pr(X = x|Y = i) = \sum_{i \in \{1,2,\ldots,K\}} \lambda_i g_i(x)$$

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

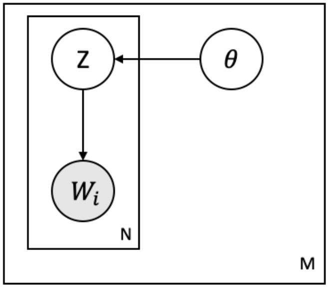
Model each *document* as having a single hidden topic.



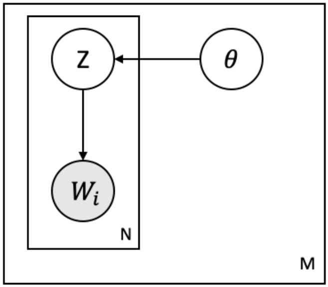
Each topic defines a distribution over words.



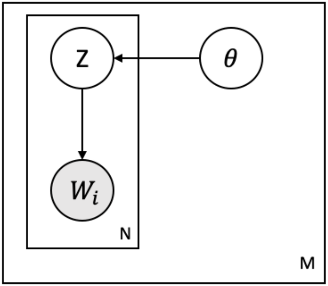
Model each *word* as having a single hidden topic.



Each document has a *distribution* over topics.



Some topics can be shared across all documents.



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

$$\Pr(w_1, w_2, \ldots, w_N | z) = \prod_{i=1}^{N} \Pr(w_i | z)$$

How can the topic variable help here?

- 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:

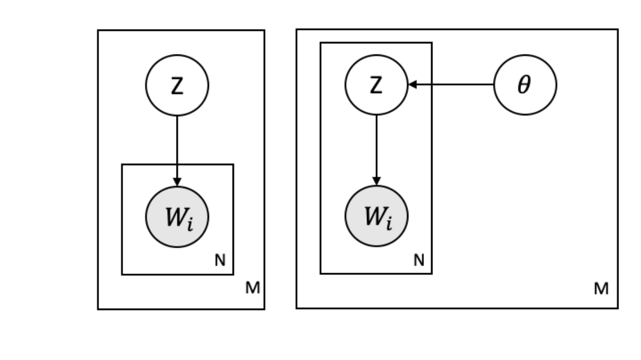
- Generate a topic Z for a word

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

- Generate a word W according to the topic distribution

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

MIXTURE VS. LDA MODEL



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*.

- Two missing variables: labels and parameters (Z, θ)

- If we knew Z, we could use MLE to estimate θ

- If we knew θ, we could use Bayes' rule to infer Z

- Initialize the parameters θ_0 somehow (randomly?)

- 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.

- 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.
- M-step: Update the parameters θ_{i+1} to maximize the expected log-likelihood.
- Iterate until the expected log-likelihood stops increasing.

- 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.
- 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.

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

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

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.

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

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

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

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

- Initialize the parameters θ_0 somehow (randomly?)

- 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.

- 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.
- M-step: Update the parameters θ_{i+1} to maximize the expected log-likelihood.
- Iterate until the expected log-likelihood stops increasing.

- 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.
- 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.

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

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

EM MAXIMIZES A BOUND ON THE OBSERVED LIKELIHOOD

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

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

TOPIC MIXTURE MODEL PARAMETER ESTIMATION

1. Estimate the posterior probability over topics for each document (E-step)
2. Update topic distributions using these posterior probabilities as fractional counts (M-step)

Given topic priors $p_i = \Pr(Z = i)$ and topic conditional probabilities $t_{ij} = (W = j|Z = i)$

$$\Pr(Z = i|w_1, w_2, \ldots) = \frac{p_i \prod_j t_{ij}}{\sum_k p_k \prod_n t_{kj}}$$

[K-means]

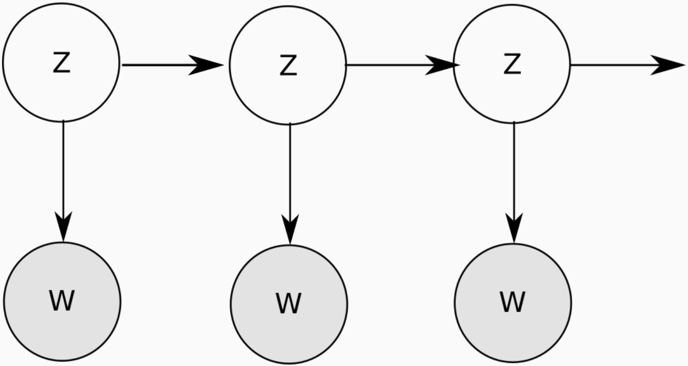
Assign document to topic with highest posterior (aka hard EM)

$$i^* = \underset{i}{\operatorname{argmax}} \Pr(Z = i | w_1, w_2, \ldots)$$

[Gibbs sampling]

Sample a topic from the posterior

$$i^* \sim \Pr(Z = i | w_1, w_2, \ldots)$$



Useful for tagging, segmentation, speech, etc.

HIDDEN MARKOV MODEL TASKS

- How might a 2-state HMM model English text?

- How could we use an HMM to solve a substitution cipher?

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:

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

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:

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

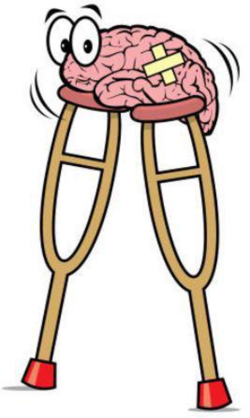
Posterior probabilities over single states

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

Posterior probabilities over state transitions

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

HACK OF THE DAY



Initialize complex models with parameters from simpler ones.