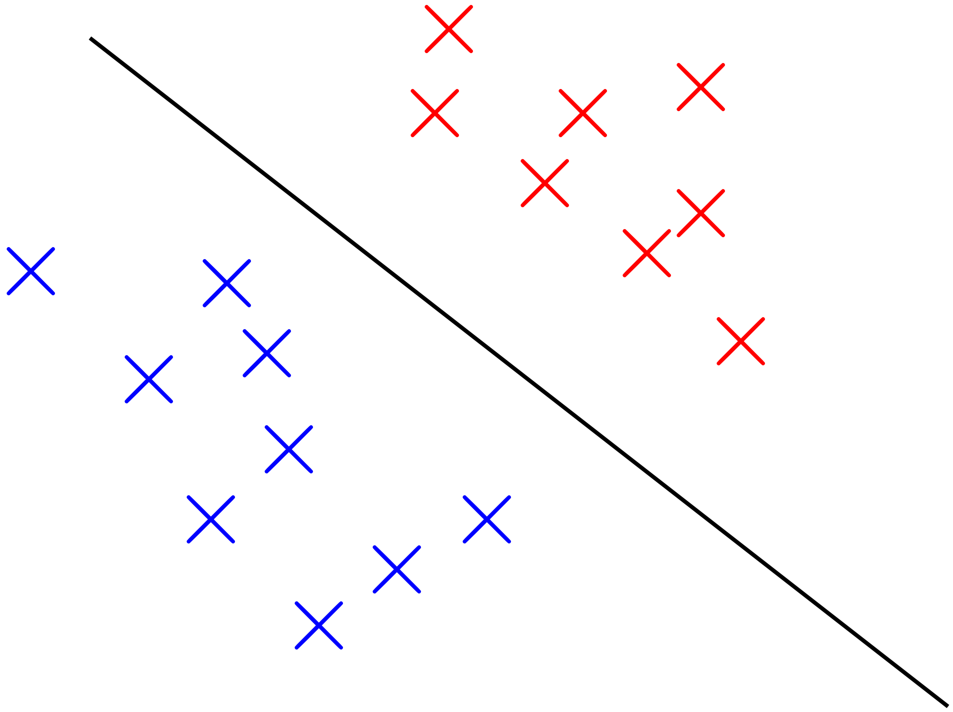
50.007
Machine Learning

Lu, Wei

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

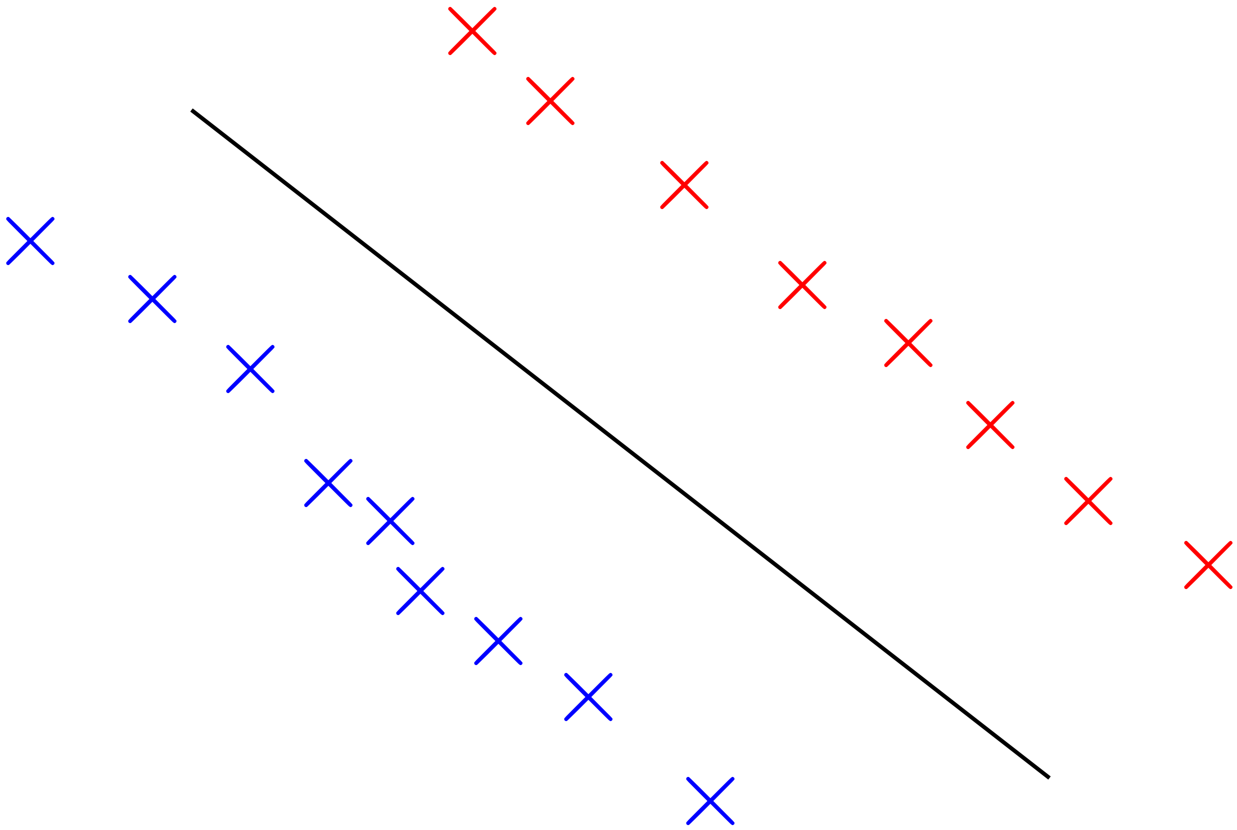
Generative Models, Naive Bayes

Linear Classification



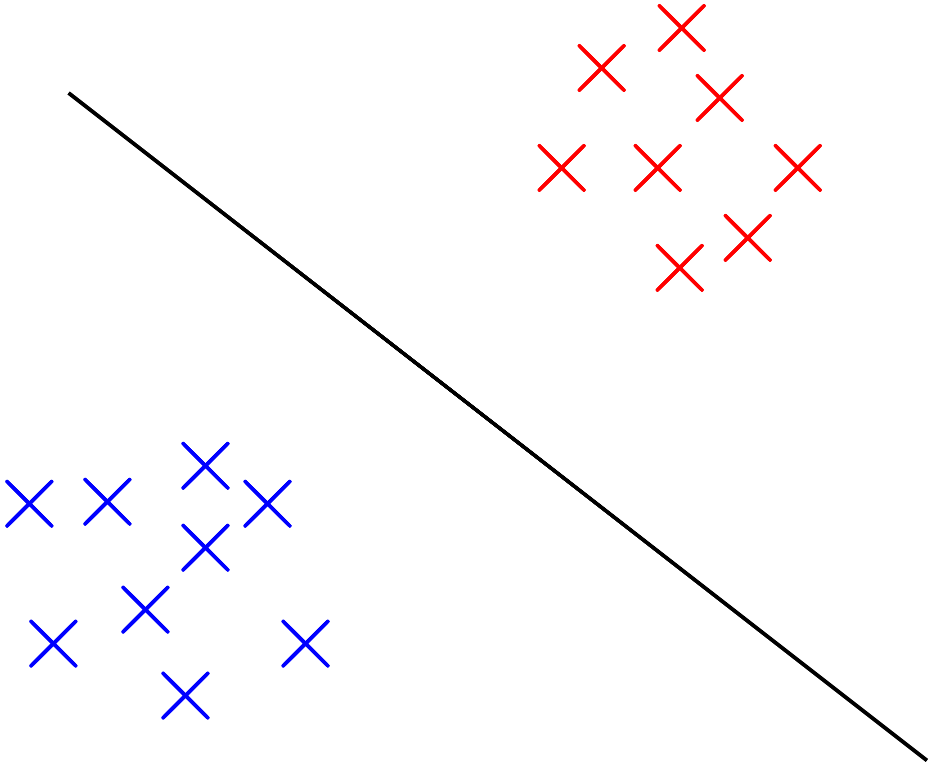
A linear decision boundary

Linear Classification



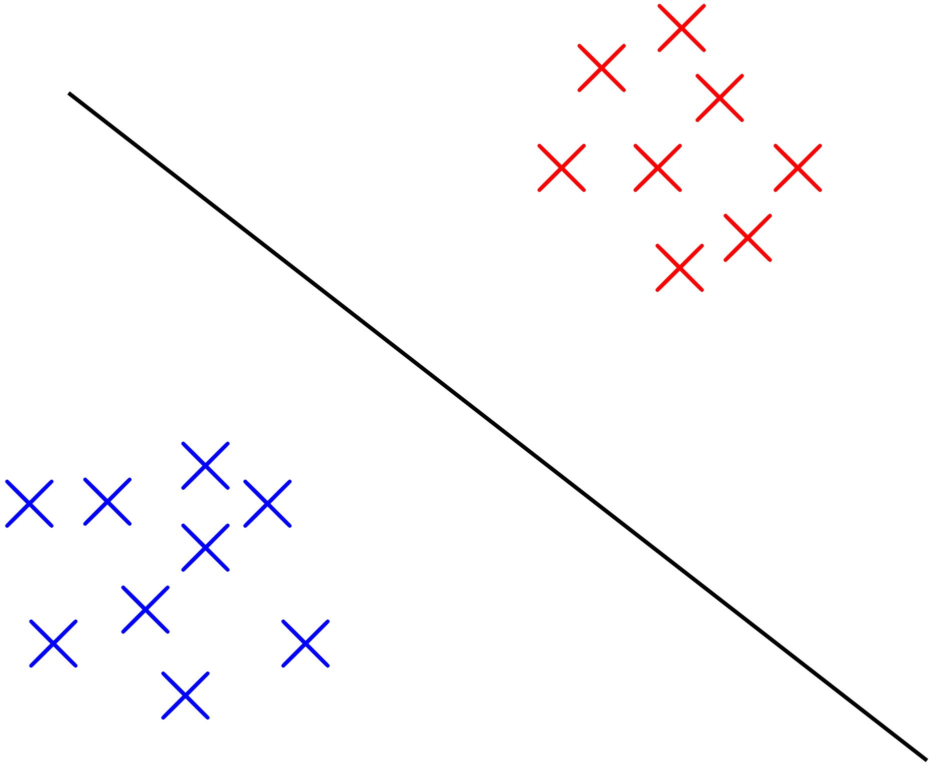
The same linear decision boundary

Linear Classification



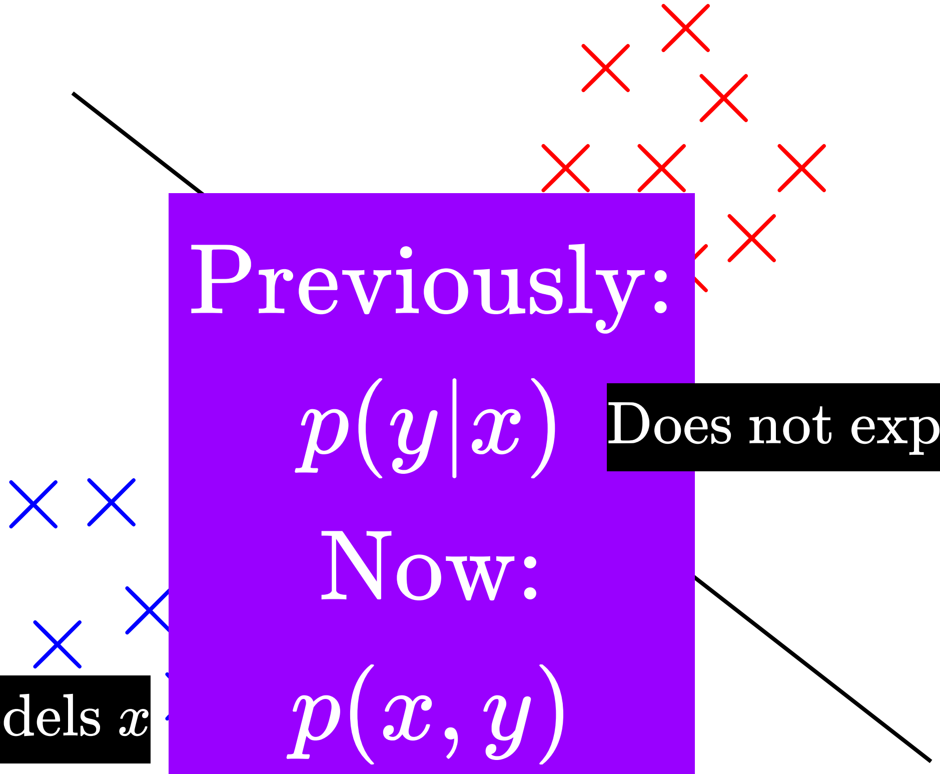
The same linear decision boundary

Linear Classification



Can we have another model that is able to say/capture something about the inputs?

Linear Classification



Previously:

$p(y|x)$

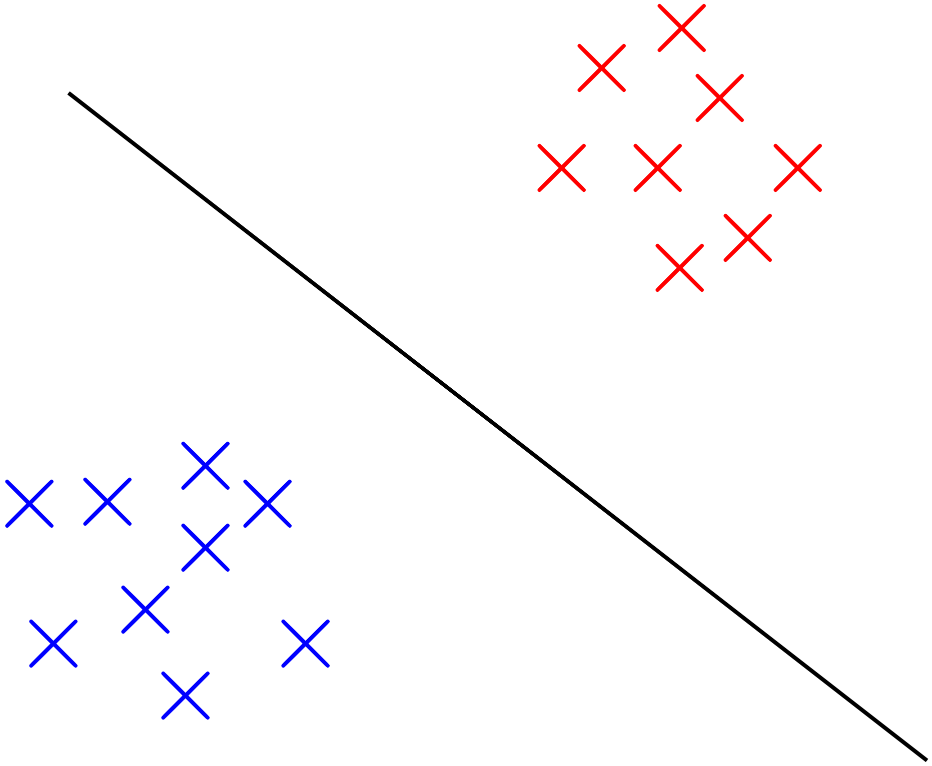
Does not explicitly model x

Now:

$p(x, y)$

Explicitly models x

Linear Classification

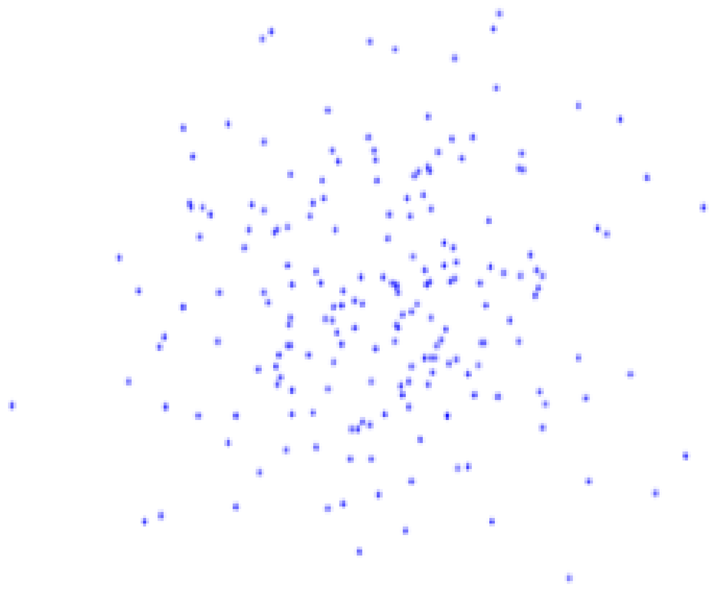


$$p(x, y)$$

We will assume there is an underlying "generative process" for both input and output!

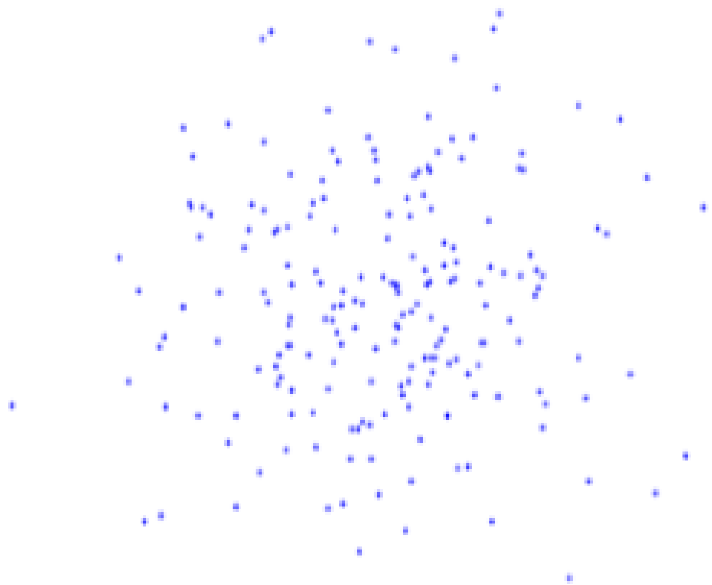
Generative Process

Gaussian



Each point is generated from the same underlying Gaussian distribution.

Generative Process
Gaussian



The likelihood for each point:

$$p(x; \mu, \sigma) = C \cdot e^{-\frac{1}{2\sigma^2}||x-\mu||^2}$$

Point

Mean

Standard Deviation

Generative Process

Multinomial

$$a \; b \; c \; d$$

What is the probability of "generating" this document with the above 4 words?

$$p(a, b, c, d)$$

Generative Process

Multinomial

$$a$$

What is the probability of "generating" this document with the above 4 words?

$$p(a, b, c, d)$$

$$p(a)$$

Generative Process
Multinomial

$$a\ b$$

What is the probability of "generating" this document with the above 4 words?

$$p(a, b, c, d)$$

$$p(a) \times p(b|a)$$

Generative Process
Multinomial

$$a\ b\ c$$

What is the probability of "generating" this document with the above 4 words?

$$p(a, b, c, d)$$

$$p(a) \times p(b|a) \times p(c|a, b)$$

Generative Process

Multinomial

$$a\ b\ c\ d$$

What is the probability of "generating" this document with the above 4 words?

$$p(a, b, c, d)$$

$$p(a) \times p(b|a) \times p(c|a, b) \times p(d|a, b, c)$$

Generative Process
Multinomial

$$a\ b\ c\ d$$

What is the problem with such an assumption on generating this document?

$$p(a, b, c, d)$$

Model Parameters

$$p(a) \times p(b|a) \times p(c|a, b) \times p(d|a, b, c)$$

Generative Process
Multinomial

$$a\ b\ c\ d$$

What is the problem with such an assumption on generating this document?

$$p(a, b, c, d)$$

Model Parameters

$$p(a) \times p(b|a) \times p(c|a, b) \times p(d|a, b, c)$$

Generative Process
Multinomial

$$a\ b\ c\ d$$

There is now a multinomial distribution!

$$p(a, b, c, d) = p(a) \times p(b) \times p(c) \times p(d)$$

Model Parameters

Generative Process
Multinomial

$$a\ a\ b$$

There is now a multinomial distribution!

$$p(a, a, b) = p(a) \times p(a) \times p(b)$$

Model Parameters

Generative Process

$$D$$

Assume we have a vocabulary
$$V = \{w_1, w_2, \ldots, w_{|V|}\}$$

$$p(D) = ?$$

How do we rewrite the above probability in terms of w's?

Generative Process

$$D$$

Assume we have a vocabulary
$V = \{w_1, w_2, \dots, w_{|V|}\}$

$$p(D) = p(w_1)^{\text{count}(w_1, D)} \times p(w_2)^{\text{count}(w_2, D)} \times \cdots \times p(w_{|V|})^{\text{count}(w_{|V|}, D)}$$

The number of times you see this word $w_{|V|}$ in the document D

Generative Process

$$D$$

Assume we have a vocabulary
$V = \{w_1, w_2, \ldots, w_{|V|}\}$

$$p(D) = \prod_j p(w_j)^{\text{count}(w_j, D)}$$

The number of times you see this word w_j in the document D.

Generative Process

$D_1 \quad \ldots \quad D_i \quad \ldots \quad D_n$

Assume we have a vocabulary

$$V = \{w_1, w_2, \ldots, w_{|V|}\}$$

$$p(D_i) = \prod_j p(w_j)^{\text{count}(w_j, D_i)}$$

What is the overall objective defined over the entire training set?

Generative Process

D_1 $\quad \dots \quad$ D_i $\quad \dots \quad$ D_n

Assume we have a vocabulary
$V = \{w_1, w_2, \dots, w_{|V|}\}$

$$p(D_i) = \prod_j p(w_j)^{\text{count}(w_j, D_i)}$$

The overall objective defined at the entire training set:

$$\prod_i p(D_i) = \prod_i \prod_j p(w_j)^{\text{count}(w_j, D_i)}$$

Generative Process

$$\prod_i p(D_i) = \prod_{i=1}^{n} \prod_{j=1}^{|V|} p(w_j)^{\text{count}(w_j, D_i)}$$

$$= \prod_{j=1}^{|V|} p(w_j)^{\text{count}(w_j, \mathcal{D})}$$

The number of times you see this word w_j in the training set $\mathcal{D}$.

Generative Process

$$\max \quad \prod_{j=1}^{|V|} p(w_j)^{\mathrm{count}(w_j,\mathcal{D})}$$

$$\min \quad -\log \prod_{j=1}^{|V|} p(w_j)^{\mathrm{count}(w_j,\mathcal{D})}$$

$$\min \quad -\sum_{j=1}^{|V|} \log p(w_j)^{\mathrm{count}(w_j,\mathcal{D})}$$

$$\min \quad -\sum_{j=1}^{|V|} \mathrm{count}(w_j,\mathcal{D}) \times \log p(w_j)$$

We arrived at this minimization problem now.
What shall we do next?

Generative Process

$$\max \quad \prod_{j=1}^{|V|} p(w_j)^{\mathrm{count}(w_j, \mathcal{D})}$$

$$\min \quad -\log \prod_{j=1}^{|V|} p(w_j)^{\mathrm{count}(w_j, \mathcal{D})}$$

$$\min \quad -\sum_{j=1}^{|V|} \log p(w_j)^{\mathrm{count}(w_j, \mathcal{D})}$$

$$\min \quad \underbrace{-\sum_{j=1}^{|V|} \mathrm{count}(w_j, \mathcal{D}) \times \log p(w_j)}_{\ell}$$

$$\frac{\partial \ell}{\partial p(w_j)}?$$

Generative Process

$$\min \quad -\sum_{j=1}^{|V|} \text{count}(w_j, \mathcal{D}) \times \log p(w_j)$$

One constraint: $\sum_j p(w_j) = 1$

$$\min \quad -\sum_{j=1}^{|V|-1} \text{count}(w_j, \mathcal{D}) \times \log p(w_j)$$

$$-\text{count}(w_{|V|}, \mathcal{D}) \times \log \Big(1 - \underbrace{\sum_{k=1}^{|V|-1} p(w_k)}_{\text{contains } w_j}\Big)$$

$p(w_{|V|}) = 1 - \sum_{k=1}^{|V|-1} p(w_k)$

Generative Process

$$\min \quad -\sum_{j=1}^{|V|-1} \text{count}(w_j, \mathcal{D}) \times \log p(w_j)$$
$$-\text{count}(w_{|V|}, \mathcal{D}) \times \log\left(1 - \sum_{k=1}^{|V|-1} p(w_k)\right)$$

$$\frac{\partial \ell}{\partial p(w_j)} = -\frac{\text{count}(w_j, \mathcal{D})}{p(w_j)} + \frac{\text{count}(w_{|V|}, \mathcal{D})}{p(w_{|V|})}$$

Generative Process

$$\min \quad -\sum_{j=1}^{|V|-1} \text{count}(w_j, \mathcal{D}) \times \log p(w_j)$$

$$-\text{count}(w_{|V|}, \mathcal{D}) \times \log \left(1 - \sum_{k=1}^{|V|-1} p(w_k)\right)$$

$$\frac{\partial \ell}{\partial p(w_j)} = -\frac{\text{count}(w_j, \mathcal{D})}{p(w_j)} + \frac{\text{count}(w_{|V|}, \mathcal{D})}{p(w_{|V|})} = 0$$

Generative Process

$$\min \quad -\sum_{j=1}^{|V|-1} \text{count}(w_j, \mathcal{D}) \times \log p(w_j)$$
$$-\text{count}(w_{|V|}, \mathcal{D}) \times \log\left(1 - \sum_{k=1}^{|V|-1} p(w_k)\right)$$

$$\frac{\partial \ell}{\partial p(w_j)} = -\frac{\text{count}(w_j, \mathcal{D})}{p(w_j)} + \frac{\text{count}(w_{|V|}, \mathcal{D})}{p(w_{|V|})} = 0$$

$$\frac{\text{count}(w_j, \mathcal{D})}{p(w_j)} = \frac{\text{count}(w_{|V|}, \mathcal{D})}{p(w_{|V|})}$$

Generative Process

$$\min \quad -\sum_{j=1}^{|V|-1} \text{count}(w_j, \mathcal{D}) \times \log p(w_j)$$

$$-\text{count}(w_{|V|}, \mathcal{D}) \times \log\left(1 - \sum_{k=1}^{|V|-1} p(w_k)\right)$$

$$\frac{\partial \ell}{\partial p(w_j)} = -\frac{\text{count}(w_j, \mathcal{D})}{p(w_j)} + \frac{\text{count}(w_{|V|}, \mathcal{D})}{p(w_{|V|})} = 0$$

$$\frac{\text{count}(w_j, \mathcal{D})}{p(w_j)} = \frac{\text{count}(w_{|V|}, \mathcal{D})}{p(w_{|V|})}$$

$$\frac{\text{count}(w_j, \mathcal{D})}{p(w_j)} = \frac{\sum_k \text{count}(w_k, \mathcal{D})}{\sum_k p(w_k)} = \frac{\sum_k \text{count}(w_k, \mathcal{D})}{1}$$

Generative Process

$$\min \quad -\sum_{j=1}^{|V|-1} \mathrm{count}(w_j, \mathcal{D}) \times \log p(w_j)$$
$$-\mathrm{count}(w_{|V|}, \mathcal{D}) \times \log\left(1 - \sum_{k=1}^{|V|-1} p(w_k)\right)$$

$$\frac{\partial \ell}{\partial p(w_j)} = -\frac{\mathrm{count}(w_j, \mathcal{D})}{p(w_j)} + \frac{\mathrm{count}(w_{|V|}, \mathcal{D})}{p(w_{|V|})} = 0$$

$$\frac{\mathrm{count}(w_j, \mathcal{D})}{p(w_j)} = \frac{\mathrm{count}(w_{|V|}, \mathcal{D})}{p(w_{|V|})}$$

$$p(w_j) = \frac{\mathrm{count}(w_j, \mathcal{D})}{\sum_k^{|V|} \mathrm{count}(w_k, \mathcal{D})}$$

Generative Process

D_1 $\ldots$ D_i $\ldots$ D_n

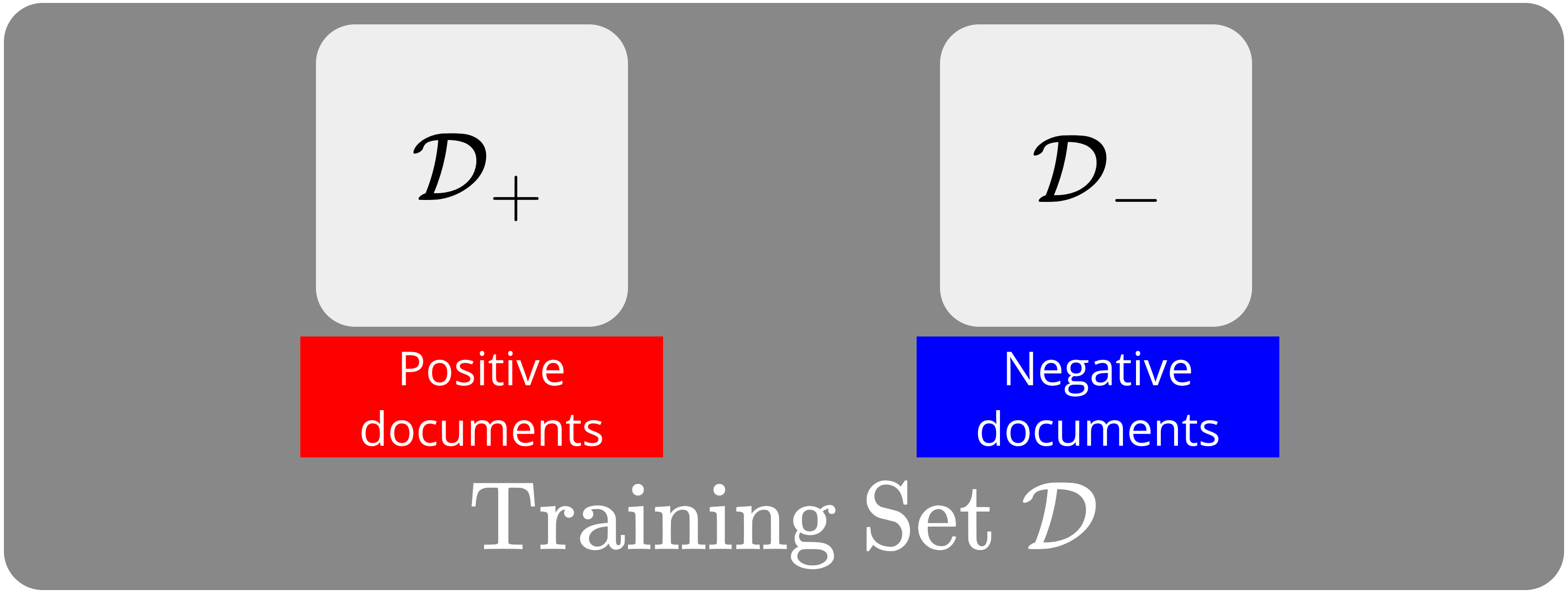
Training Set $\mathcal{D}$

The number of times we see the word w_j in the training set $\mathcal{D}$

$$p(w_j) = \frac{\text{count}(w_j, \mathcal{D})}{\sum_k^{|V|} \text{count}(w_k, \mathcal{D})}$$

The total number of words that appear in the training set $\mathcal{D}$

Naive Bayes


$\mathcal{D}_+$
$\mathcal{D}_-$
Positive documents
Negative documents
Training Set $\mathcal{D}$

Naive Bayes

$\mathcal{D}_+$

Positive documents

$\mathcal{D}_-$

Negative documents

Training Set $\mathcal{D}$

$p(\mathcal{D}_+)$ $\quad$ $p(\mathcal{D}_-)$

Naive Bayes

$\mathcal{D}_+$

Positive documents

$\mathcal{D}_-$

Negative documents

Training Set $\mathcal{D}$

$p(\mathcal{D}_+)$ $p(\mathcal{D}_-)$

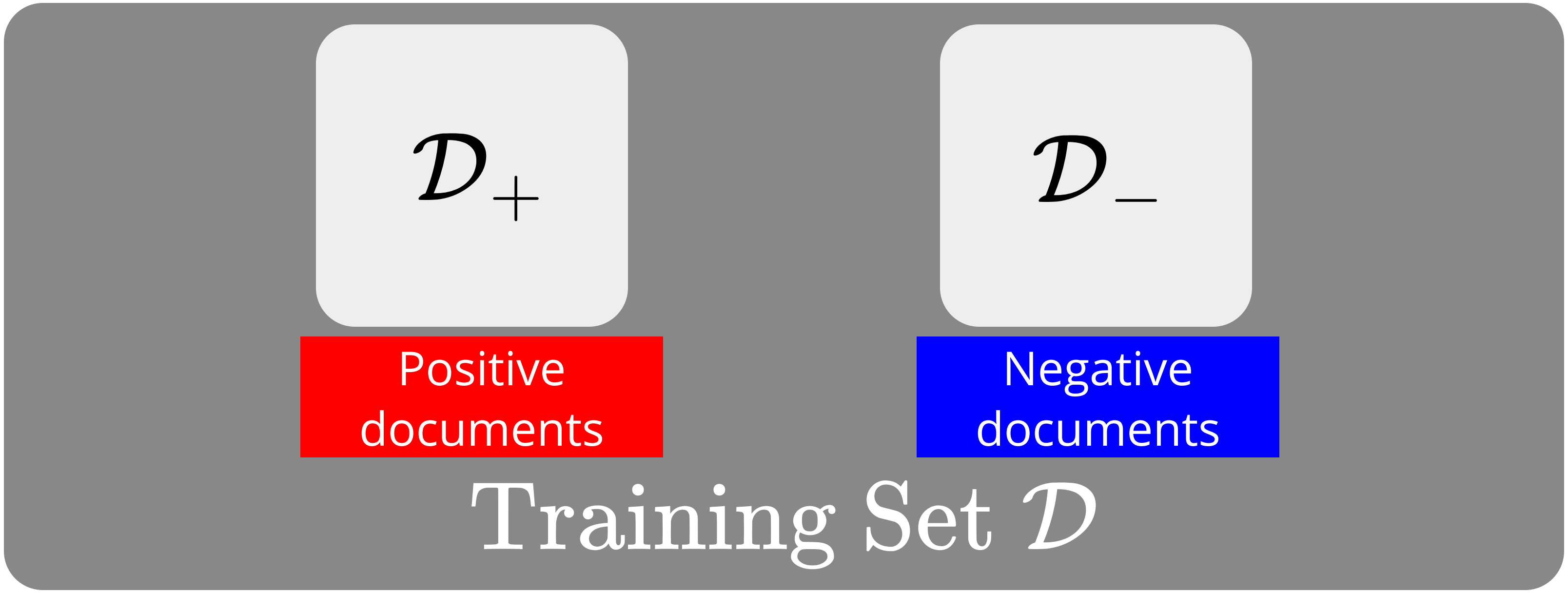
$$p(y = +1)^{\text{count}(y=+1)}$$

$$\prod_w \theta_w^{+\,n(w,+)}$$

$\text{count}(w, +)$

$p(w|y = +1)$ for positive documents

Naive Bayes



$$p(\mathcal{D}_+) \qquad\qquad p(\mathcal{D}_-)$$

$$\prod_w {\theta_w^+}^{n(w,+)} \qquad\qquad \prod_w {\theta_w^-}^{n(w,-)}$$

$$p(y=+1)^{\text{count}(y=+1)} \qquad p(y=-1)^{\text{count}(y=-1)}$$

Naive Bayes
Evaluation

Learned model parameters θ_w^+, θ_w^-, and $p(y = +1)$

D

what should be the output y label

for this new document D?

Naive Bayes

Evaluation

Learned model parameters θ_w^+, θ_w^-, and $p(y=+1)$

$$D$$

$$\frac{p(y=+1|D)}{p(y=-1|D)}$$

Naive Bayes
Evaluation

Learned model parameters θ_w^+, θ_w^-, and $p(y=+1)$

D

$$\frac{p(D|\theta^+) \times p(y=+1)}{p(D|\theta^-) \times p(y=-1)}$$

Naive Bayes

Evaluation

Learned model parameters θ_w^+, θ_w^-, and $p(y = +1)$

D

$$\log \frac{p(D|\theta^+) \times p(y = +1)}{p(D|\theta^-) \times p(y = -1)}$$

Naive Bayes

Evaluation

Learned model parameters θ_w^+, θ_w^-, and $p(y = +1)$

D

$$\log \frac{P(D|\theta^+)}{P(D|\theta^-)} = \log P(D|\theta^+) - \log P(D|\theta^-)$$

$$= \sum_w n(w)(\log \theta_w^+ - \log \theta_w^-)$$

$$= \sum_w n(w) \log \underbrace{\frac{\theta_w^+}{\theta_w^-}}_{=\theta_w}$$

Naive Bayes
Evaluation

Learned model parameters θ_w^+, θ_w^-, and $p(y = +1)$

D

$$\log \frac{P(D|\theta^+)}{P(D|\theta^-)} = \Phi(D) \cdot \theta = \begin{bmatrix} n(w_1) \\ \vdots \\ n(w_{|V|}) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_{|V|} \end{bmatrix}$$

Naive Bayes
Evaluation

Learned model parameters θ_w^+, θ_w^-, and $p(y=+1)$

D

$$\log \frac{P(D|\theta^+)P(y{=}+)}{P(D|\theta^-)P(y{=}-)} = \sum_w n(w) \underbrace{\log \frac{\theta_w^+}{\theta_w^-}}_{\theta_w} + \underbrace{\log \frac{P(y=+)}{P(y=-)}}_{=\theta_0}$$

$$= \sum_w n(w)\theta_w + \theta_0$$

Question

What about a new word that did not appear in the training set?

Question

What about a new word that did not appear in the training set?

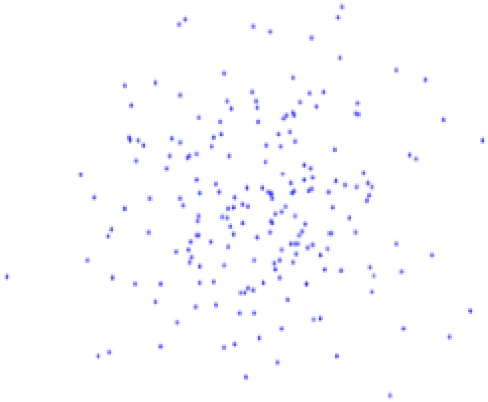
Ignore that word.

Question

What if a word only appears in the positive documents in the training set?

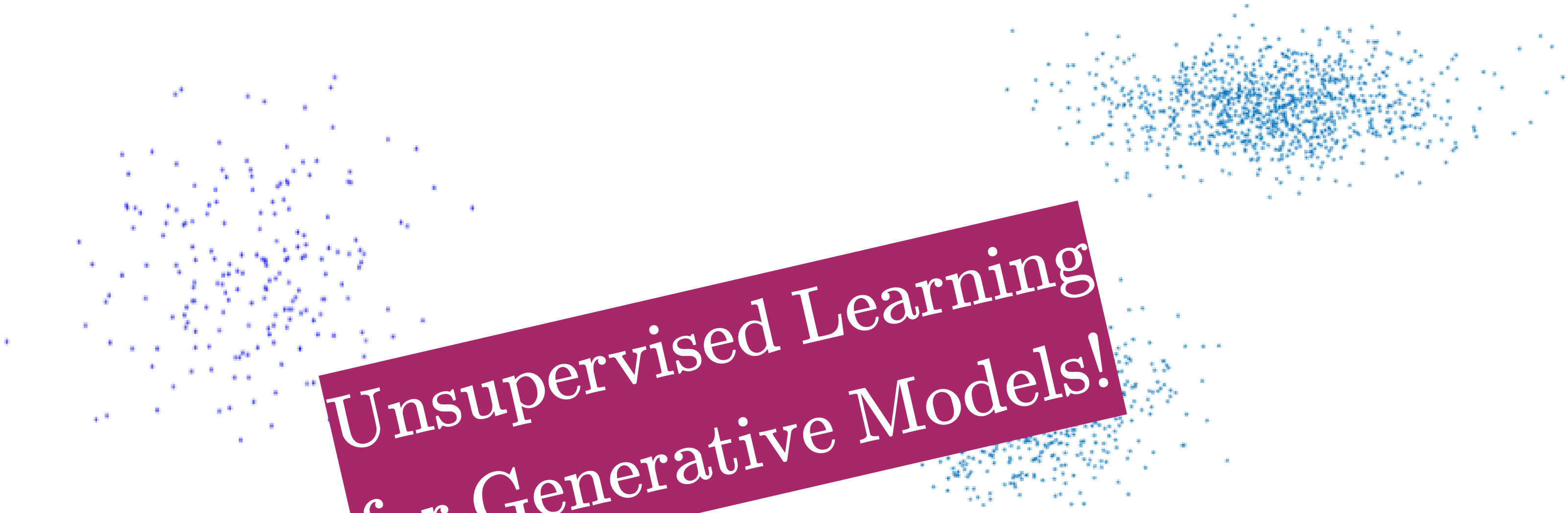
Use smoothing. Assume there is a small amount of times for each word to appear in positive/negative documents, as long as it appears in the training set.

Generative Process



What if the data looks like this?

Generative Process



Unsupervised Learning for Generative Models!

What if the data looks like this?

Generative Process

Learn Mixture Model with Expectation Maximization

What if the data looks like this?