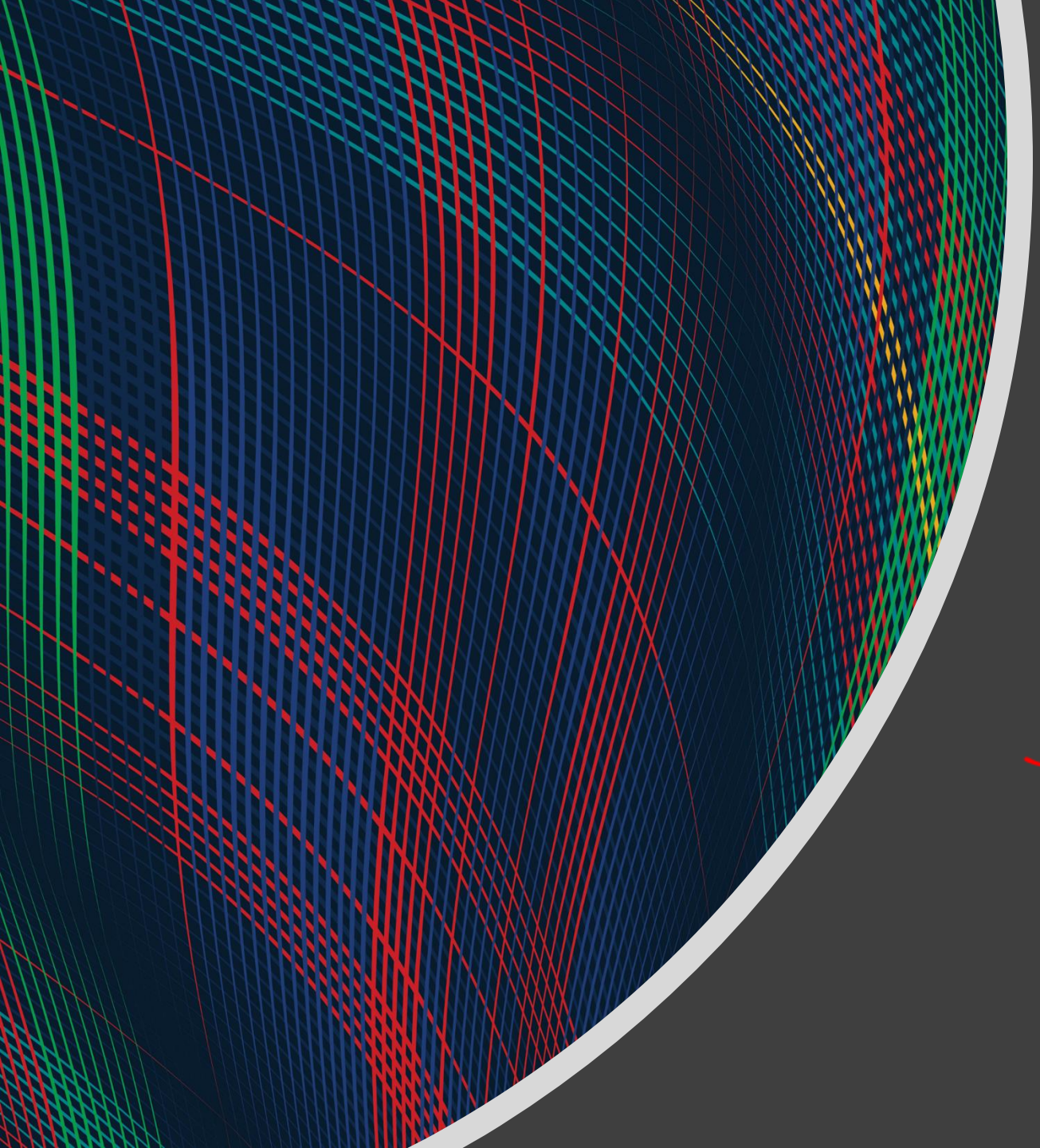


10-315 Introduction to ML

Probabilistic Models: Generative and Naïve Bayes

Instructor: Pat Virtue

Poll 1

Let's say we gave an exam that had 3 different versions where the probability of different exams was modeled as a **categorical** random variable Y and the scores from each exam were modeled as $X_{Y=k} \sim \mathcal{N}(\mu_k, \sigma_k^2)$. $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$

How many parameters are in this **generative** model?

A. 2

B. 3

C. 6

D. 9

E. 3N

$$p(x, y) = p(x \mid y)\, p(y)$$

$$p(y \mid x) = \frac{p(x \mid y) p(y)}{\sum_{y=k}^{3} p(x \mid y) p(y)}$$

$\mathcal{N}(\mu_1, \sigma_1)$ ϕ_1

$\mathcal{N}(\mu_2, \sigma_2)$ ϕ_2

$\mathcal{N}(\mu_3, \sigma_2)$ ϕ_3

Poll 2

Let's say we gave an exam that had 3 different versions where the probability of different exams Y given the scores from the exam x was modeled as a **multiclass logistic regression** $p(Y_k = 1 \mid x)$ trained on dataset $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$

$x \in \mathbb{R}$

How many parameters are in this **discriminative** model?

A. 2

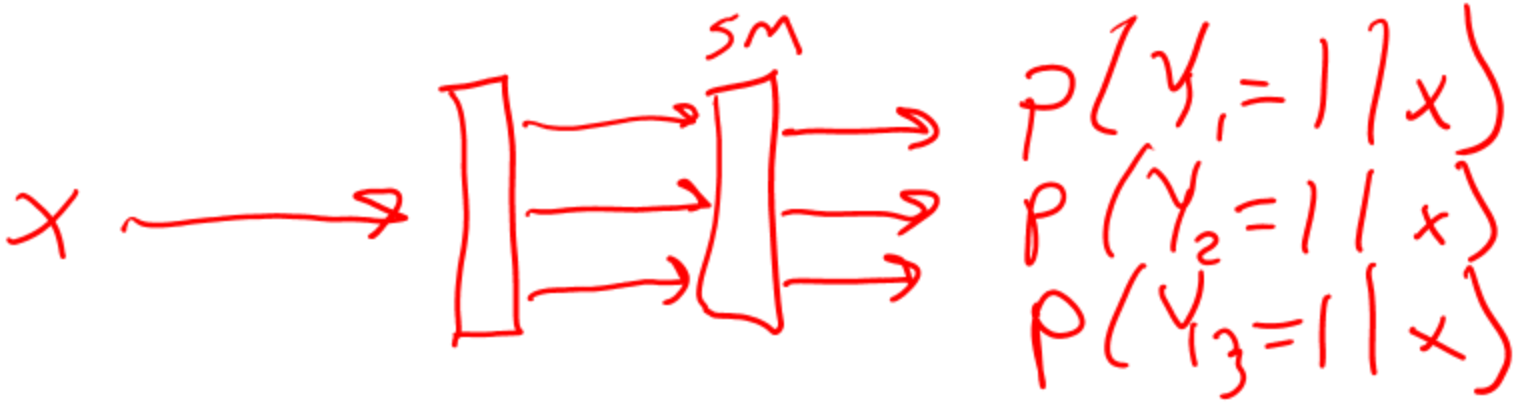
B. 3

C. 6

D. 9

E. 3N

$p(Y_k \mid x) = g_{softmax}(Wx + b)$



Poll 3

Which of the following probability models will allow us to sample new data, $\mathcal{D} = \left\{x_1^{(i)}, x_2^{(i)}, y^{(i)}\right\}_{i=1}^{N}$ where x_1, x_2, y are length, weight, and species (cat or dog), respectively? Assume that we have learned all of the corresponding parameters

A) Logistic regression

$$p(Y = cat \mid x_1, x_2, w_1, w_2, b)$$

B) Bernoulli + Gaussian

$$p(Y = cat \mid \phi)$$

$$p(\, x_1 \mid cat, \mu_{cat,1}, \sigma_{cat,1}\,)$$

$$p(\, x_2 \mid cat, \mu_{cat,2}, \sigma_{cat,2}\,)$$

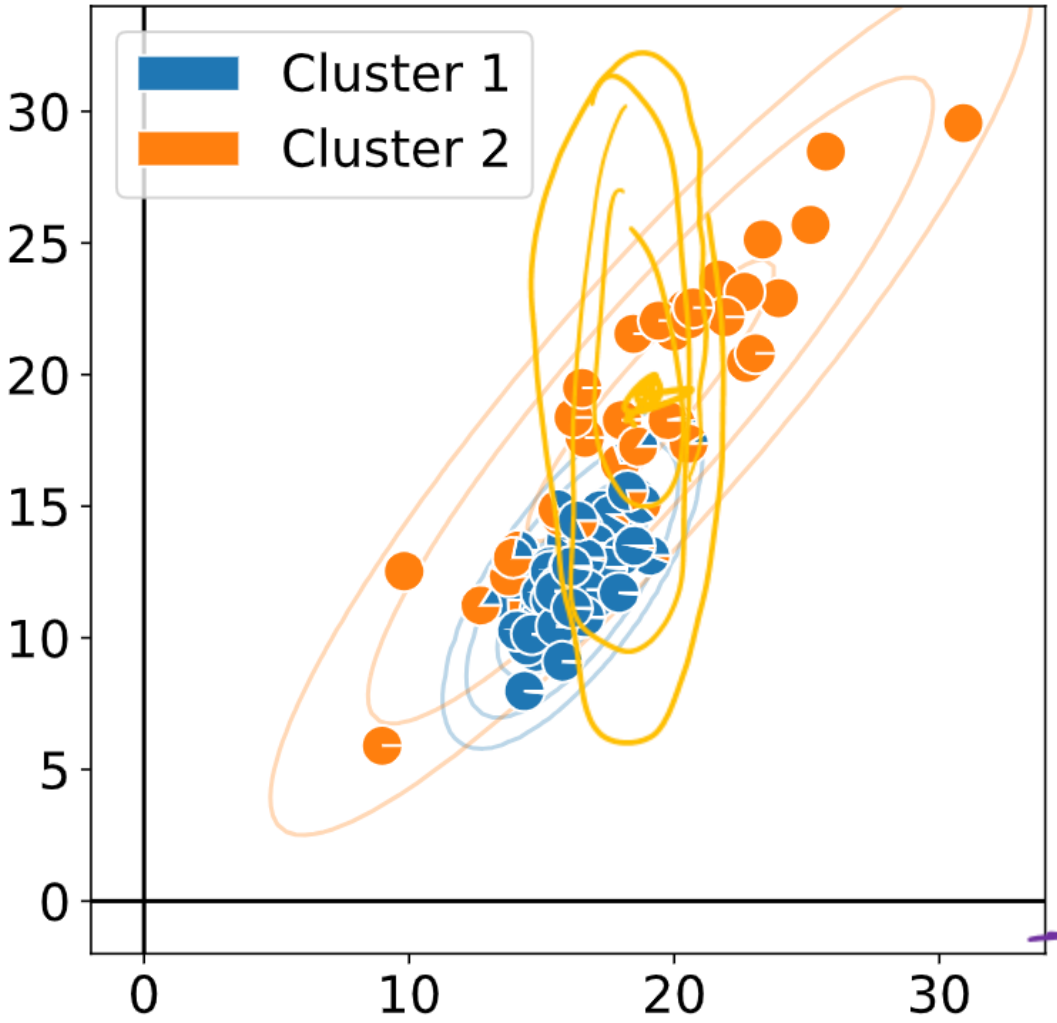
$$p(\, x_1 \mid dog, \mu_{dog,1}, \sigma_{dog,1}\,)$$

$$p(\, x_2 \mid dog, \mu_{dog,2}, \sigma_{dog,2}\,)$$

C) None of the above

Sampling Cats and Dogs

$p(x|y)p(y)$



y, x_1, x_2

$p(x_1, x_2 \mid y=dog) = N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix}\right)$

B) Bernoulli + Gaussian

Naive Bayes assumption

$p(Y = cat \mid \phi)$ ← dog

$p(x_1 \mid cat, \mu_{cat,1}, \sigma_{cat,1})$

$p(x_2 \mid cat, \mu_{cat,2}, \sigma_{cat,2})$

$p(x_1 \mid dog, \mu_{dog,1}, \sigma_{dog,1})$ ← N

$p(x_2 \mid dog, \mu_{dog,2}, \sigma_{dog,2})$ ← N

Generative Stories

Generative vs Discriminative Modeling

Discriminative: $p(y \mid x)$

Generative: $p(y \mid x) = \alpha\, p(x, y) = \alpha\, p(x \mid y)\, p(y)$

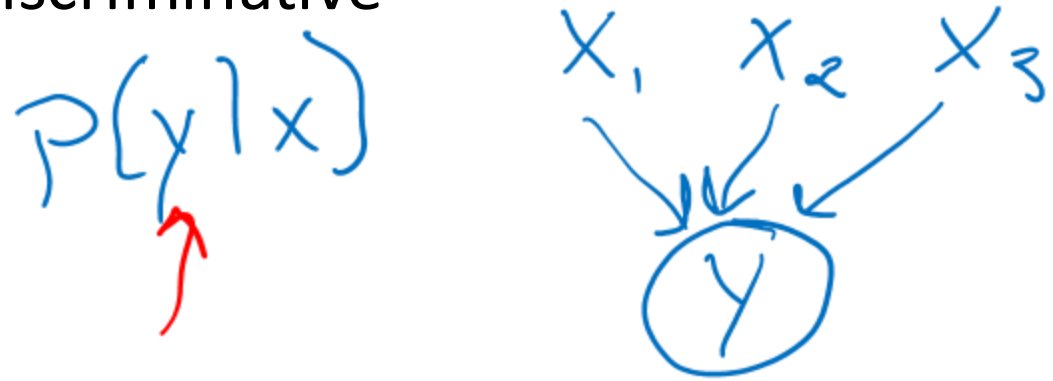
Model assumptions vs Data

- Discriminative: ↓ assump. ↑ data
- Generative: ↑ assumpt. ↓ data

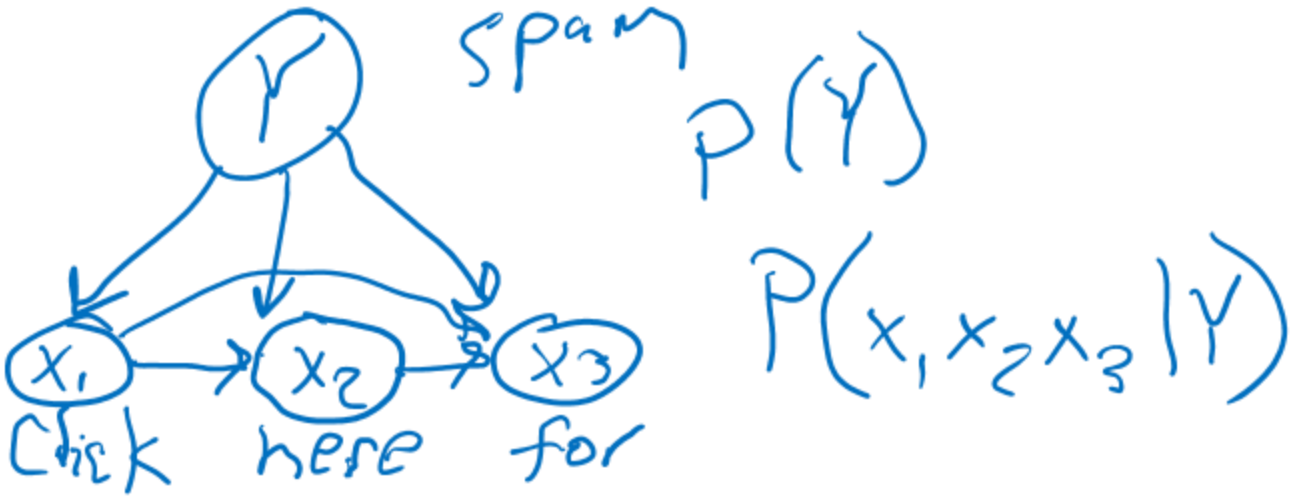
Generative Story Examples

Spam Generation

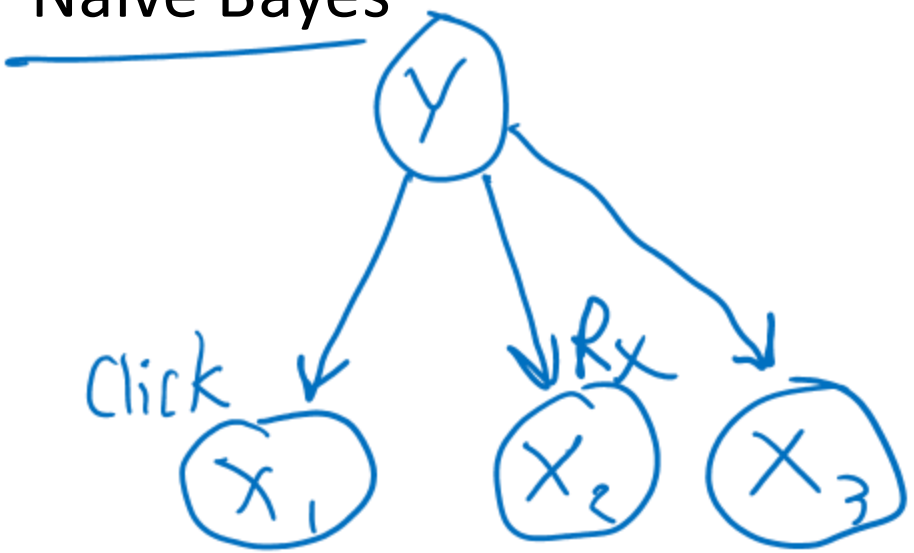
Discriminative



Generative



Generative + Naive Bayes



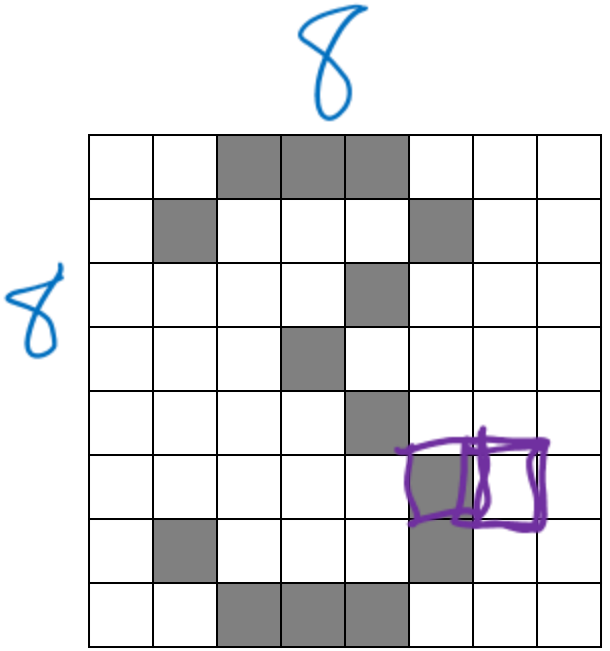
Generative Story Examples

Hand-written digits

y: Digit class: 0-9

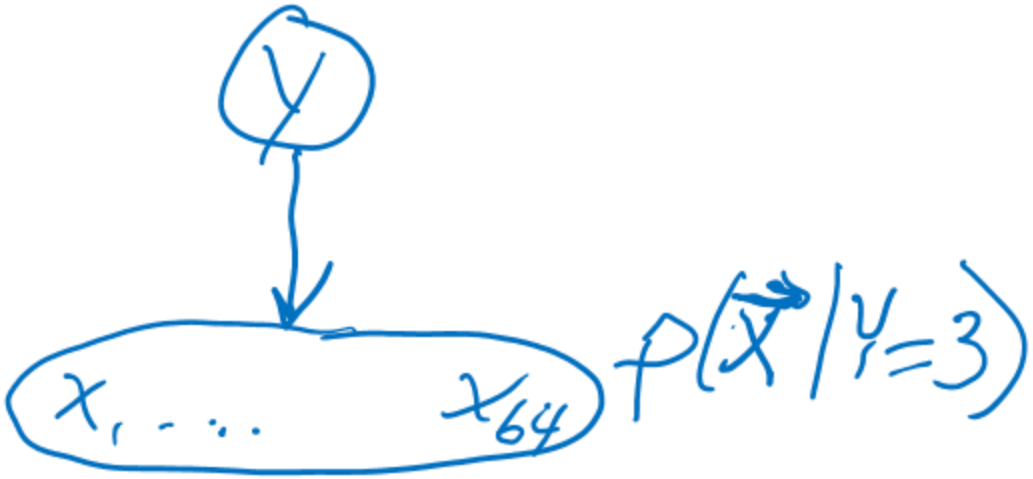
x: Pixels in images

$x_1 \dots x_{64}$

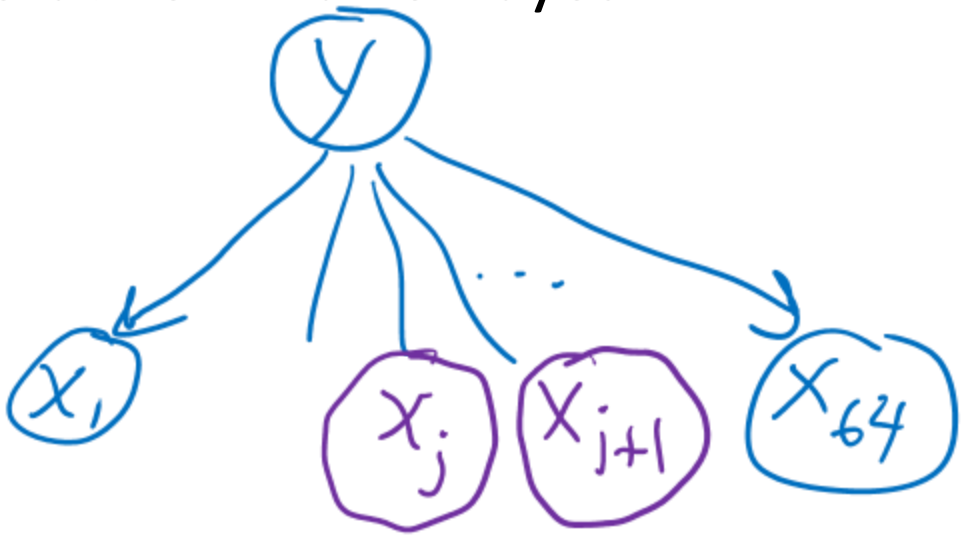


Discriminative

Generative



Generative + Naive Bayes



Discriminative and Generative

Two Applications of Bayes Rule

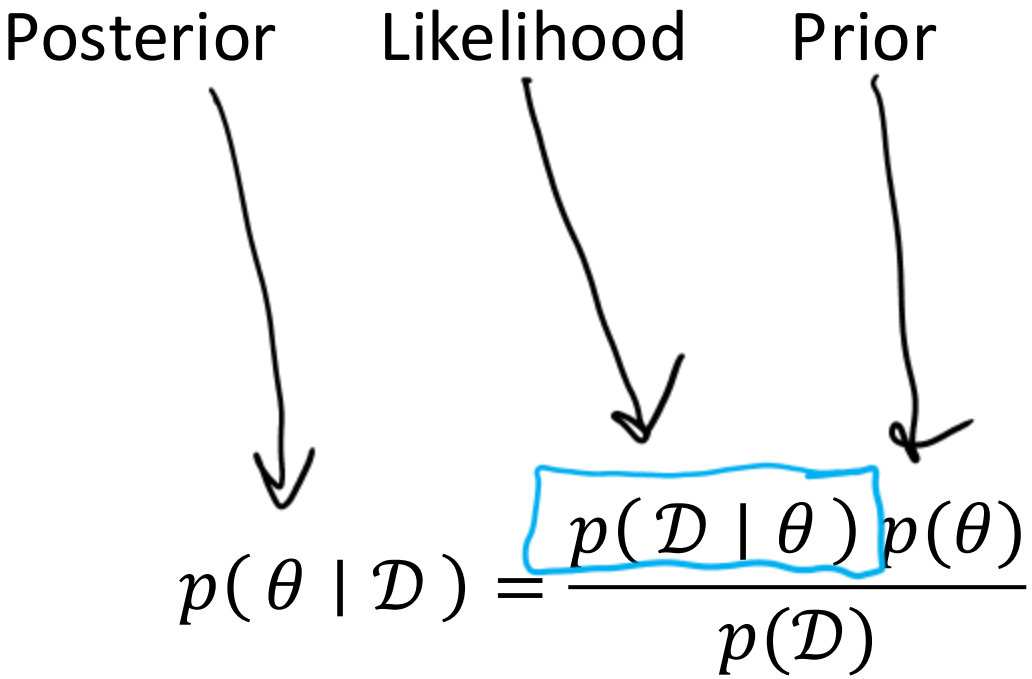
$$p(a \mid b) = \frac{p(b \mid a)\, p(a)}{p(b)}$$

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})}$$

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

Bayes Rule

Terminology



Bayes Rule

Terminology

Posterior Likelihood Prior

Conditional likelihood Posterior Class conditional Class prior

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})}$$

$$p(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

$p(x, y \mid \theta)$

$p(\mathcal{D} \mid \theta)$ $\longleftrightarrow$ $p(x, y)$ $p(x, y \mid \theta)$

Bayes Rule

Where did the parameters go?!?

$$p_\theta(y \mid x) = \frac{p(x \mid y)\, p(y)}{p(x)}$$

$$p(y \mid x, \mu_0, \mu_1, \sigma_0, \sigma_1, \phi) = \frac{p(x \mid y, \mu_y, \sigma_y)\, p(y \mid \phi)}{p(x)}$$

$$p(y \mid x, \theta) \qquad p_\theta(y \mid x) = \frac{p(x, y \mid \theta)}{p(x)}$$

Optimization: Generative vs Discriminative

Discriminative: model $p(y \mid x, \theta)$ directly

- Learn parameters θ from data

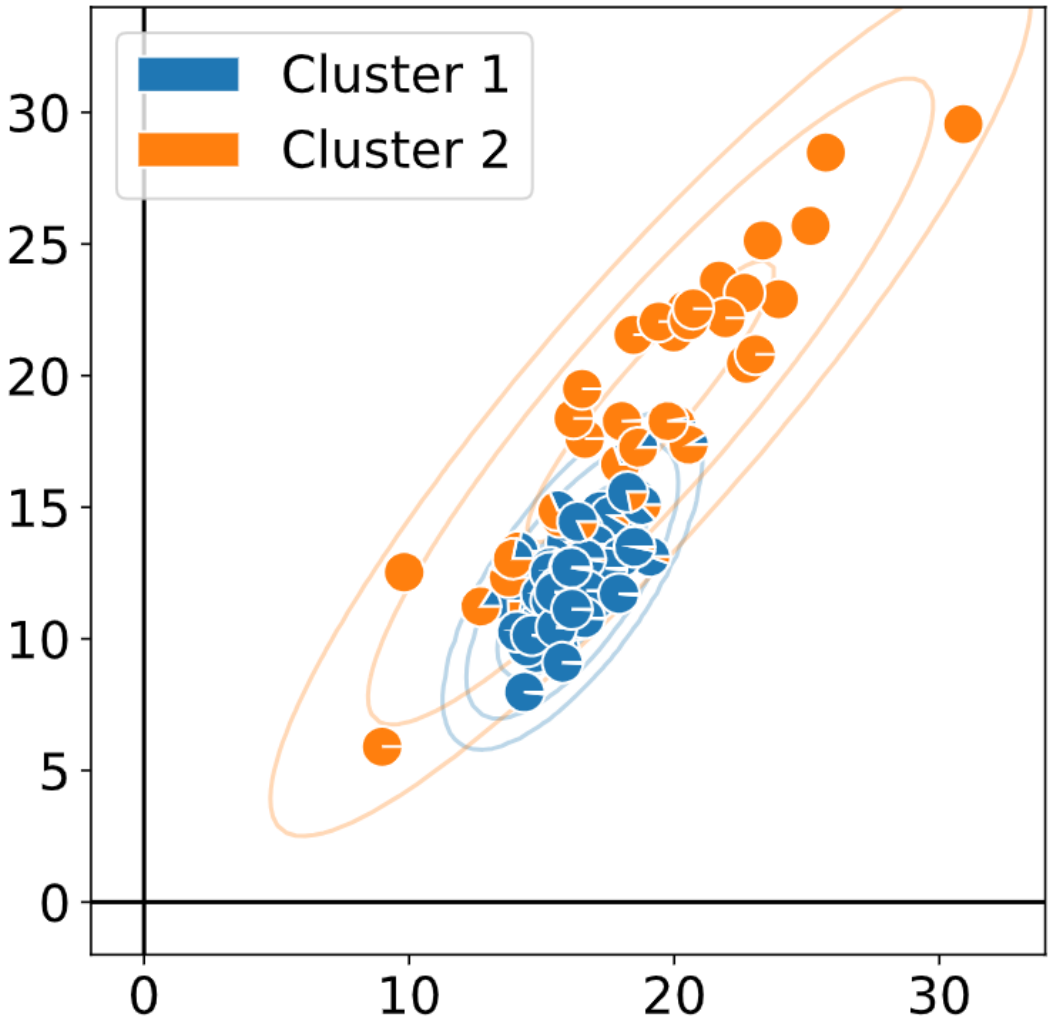
$J(\theta) = -\ell$ $\quad \frac{\partial J}{\partial \theta}$

Generative: model $p(y \mid \theta_{class})$ and $p(x \mid y, \theta_{class\ conditional})$

- Learn parameters θ_{class} and $\theta_{class\ conditional}$ from data
- Use Bayes rule to compute $p(y \mid x, \theta_{class}, \theta_{class\ conditional})$

$$p(y \mid x, \theta_{class}, \theta_{class\ conditional}) \propto p(x \mid y, \theta_{class\ conditional})\, p(y \mid \theta_{class})$$

Estimating Cats and Dogs



Bernoulli + Gaussian

$p(Y = cat \mid \phi)$ $\phi = \frac{N_{cat}}{N}$

$p(x_1, x_2 \mid Y=cat, \mu_{cat}, \Sigma_{cat})$

np.mean

np.cov

$p(x_1, x_2 \mid Y=dog, \vec{\mu}_{dog}, \Sigma_{dog})$

Reminder: Likelihood

Simple example with three i.i.d. samples of $Y \sim Bernoulli(\phi)$

$$\mathcal{D} = \{y^{(i)}\}_{i=1}^{3}$$

Likelihood

$$p(\mathcal{D}; \theta)$$

$$= p\left(y^{(1)}, y^{(2)}, y^{(3)}; \phi^{(1)}, \phi^{(2)}, \phi^{(3)}\right)$$ Expand notation

$$= p\left(y^{(1)}, y^{(2)}, y^{(3)}; \phi\right)$$ Identically distributed

$$= p\left(y^{(1)}; \phi\right) p\left(y^{(2)}; \phi\right) p\left(y^{(3)}; \phi\right)$$ Independent

$$= \prod_{i=1}^{3} p\left(y^{(i)}; \phi\right)$$

Generative MLE

$$\mathcal{D} = \{y^{(i)}, x^{(i)}\}_{i=1}^{N}$$
$$y^{(i)} \in \{0,1\}$$
$$\mathbf{x}^{(i)} \in \{0,1\}^{M}$$
$$\phi \in [0,1]$$
$$\boldsymbol{\Theta} \in [0,1]^{Mx2}$$

$$L(\phi, \boldsymbol{\Theta}) = p(\mathcal{D} \mid \phi, \boldsymbol{\Theta})$$

$$= \Pi_{n=1}^{N} p\left(\mathcal{D}^{(n)} \mid \phi, \boldsymbol{\Theta}\right) \quad \text{i.i.d assumption}$$

$$= \Pi_{n=1}^{N} p\left(\mathrm{y}^{(n)}, \mathbf{x}^{(n)} \mid \phi, \boldsymbol{\Theta}\right)$$

$$= \Pi_{n=1}^{N} p\left(\mathrm{y}^{(n)} \mid \phi\right) p\left(\mathbf{x}^{(n)} \mid \mathrm{y}^{(n)}, \boldsymbol{\Theta}\right)$$

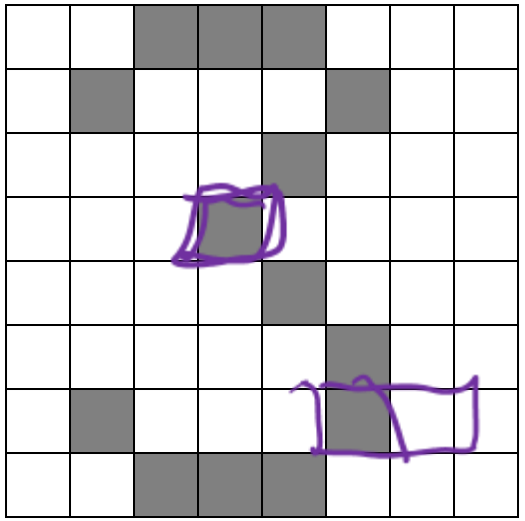
Generative model

Generative model story

$$= \Pi_{n=1}^{N} p\left(\mathrm{y}^{(n)} \mid \phi\right) p\left(x_1^{(n)}, x_2^{(n)}, \ldots, x_M^{(n)} \mid \mathrm{y}^{(n)}, \boldsymbol{\Theta}\right)$$

Naïve Bayes

Multivariate Generative Models



Hand-written digits: How many parameters?

- $P(Y)$
- $P(\boldsymbol{X} \mid Y = 3)$
  $= p(X_1, X_2, \ldots X_{64} \mid Y = 3)$

Naïve Bayes assumption (bag of pixels)

- $P(Y)$
- $P(\boldsymbol{X} \mid Y = 3)$
  $= p(X_1 \mid Y = 3)\, p(X_2 \mid Y = 3) \ldots p(X_{64} \mid Y = 3)$

Conditional Independence and Naïve Bayes

Independence

$A \perp\!\!\!\perp B$

$$P(A \mid B) = P(A)$$
$$P(B \mid A) = P(B)$$

$$P(A, B) = P(A)P(B)$$

Conditional independence

$A \perp\!\!\!\perp B \mid C$

$$P(A \mid B, C) = P(A \mid C)$$
$$P(B \mid A, C) = P(B \mid C)$$

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)$$

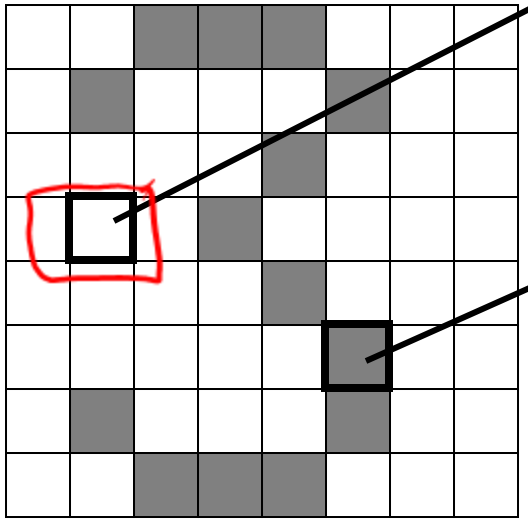
Naïve Bayes assumption

$$\forall_{j \neq k} \quad x_j \perp\!\!\!\perp x_k \mid y$$

$$p(x_1 x_2 x_3 \ldots x_M \mid y) = \prod_{j=1}^{M} p(x_j \mid y)$$

Naïve Bayes for Digits

y	$P(Y)$
1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1



y	$P(X_{3,1} = 1 \mid y)$
1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

$\phi_{y=3,x_{3,1}}$

y	$P(X_{5,5} = 1 \mid y)$
1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

SPAM Classification

Recitation Exercise

$P(Y = 0, X_1, \ldots, X_M)$	$P(Y = 1, X_1, \ldots, X_M)$

$P(Y = 0 \mid X_1, \ldots, X_M)$	$P(Y = 1 \mid X_1, \ldots, X_M)$

Naïve Bayes MLE

$$\mathcal{D} = \{y^{(n)}, \boldsymbol{x}^{(n)}\}_{n=1}^{N}$$
$$y^{(n)} \in \{0,1\}$$
$$\boldsymbol{x}^{(n)} \in \{0,1\}^{M}$$
$$\phi \in [0,1]$$
$$\boldsymbol{\Theta} \in [0,1]^{Mx2}$$

$$L(\phi, \boldsymbol{\Theta}) = p(\mathcal{D} \mid \phi, \boldsymbol{\Theta})$$

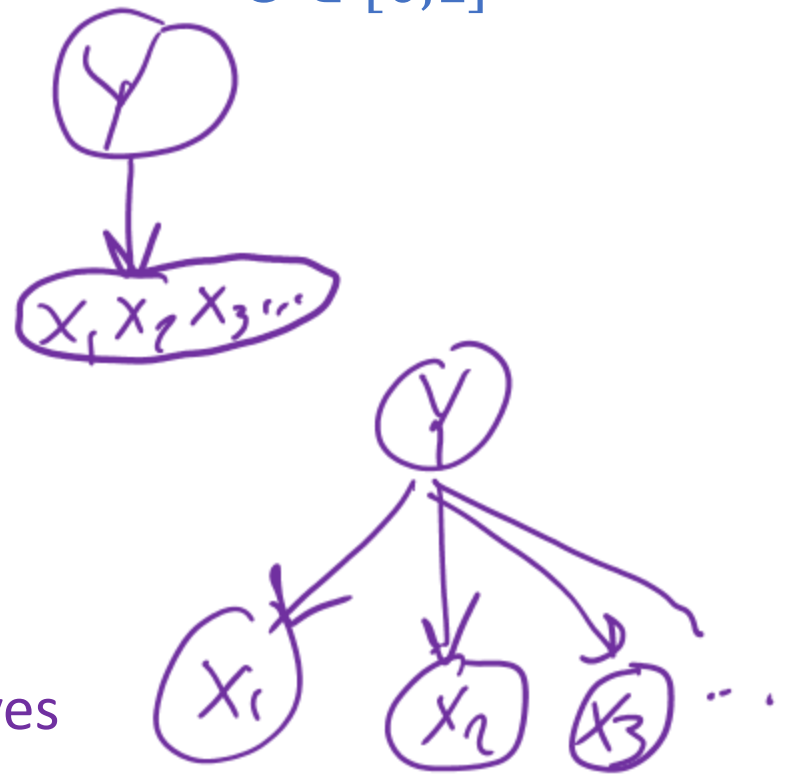
$$= \Pi_{n=1}^{N} p\left(\mathcal{D}^{(n)} \mid \phi, \boldsymbol{\Theta}\right) \quad \text{i.i.d assumption}$$

$$= \Pi_{n=1}^{N} p\left(\mathrm{y}^{(n)}, \boldsymbol{x}^{(n)} \mid \phi, \boldsymbol{\Theta}\right)$$

$$= \Pi_{n=1}^{N} p\left(\mathrm{y}^{(n)} \mid \phi\right) p\left(\boldsymbol{x}^{(n)} \mid \mathrm{y}^{(n)}, \boldsymbol{\Theta}\right) \quad \text{Generative model}$$

$$= \Pi_{n=1}^{N} p\left(\mathrm{y}^{(n)} \mid \phi\right) p\left(x_1^{(n)}, x_2^{(n)}, \ldots, x_M^{(n)} \mid \mathrm{y}^{(n)}, \boldsymbol{\Theta}\right)$$

$$= \Pi_{n=1}^{N} p\left(\mathrm{y}^{(n)} \mid \phi\right) \Pi_{m=1}^{M} p\left(x_m^{(n)} \mid \mathrm{y}^{(n)}, \theta_{m,y}\right) \quad \text{Naïve Bayes}$$



Generative Models

SPAM:

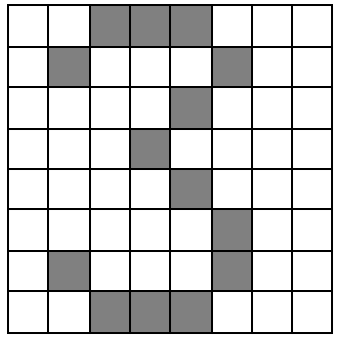
- Class distribution: $Y \sim Bern(\phi)$
- Class conditional distribution: $X_m \sim Bern(\theta_{m,y})$
- Naïve Bayes X_i conditionally independent X_j given Y for all $i \neq j$

$$p(X_i, X_j \mid Y) = p(X_i \mid Y) \mid p(X_j \mid Y)$$

Digits:

- Class distribution: $Y \sim Categorical(\boldsymbol{\phi})$
- Class conditional distribution: $X_m \sim Bern(\theta_{m,y})$
- Naïve Bayes X_i conditionally independent X_j given Y for all $i \neq j$

$$p(X_i, X_j \mid Y) = p(X_i \mid Y) \mid p(X_j \mid Y)$$



Generative Models with Continuous Features

Iris dataset:

- Class distribution: $Y \sim Bern(\phi)$
- Class conditional distribution: Multivariate Gaussian $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$
- Naïve Bayes assumption?

Poll 5

Iris dataset:

- Class distribution: $Y \sim Bern(\phi)$
- Class conditional distribution: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$
- Naïve Bayes assumption?

Which of the following pairs of Gaussian class conditional distributions satisfy the Naïve Bayes assumptions? Select ALL that apply.

A. $\boldsymbol{\mu}_{y=0} = \begin{bmatrix}-1\\0\end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \quad \boldsymbol{\mu}_{y=1} = \begin{bmatrix}1\\0\end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$

B. $\boldsymbol{\mu}_{y=0} = \begin{bmatrix}-1\\0\end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \quad \boldsymbol{\mu}_{y=1} = \begin{bmatrix}1\\0\end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix}3 & 0\\0 & 3\end{bmatrix}$

C. $\boldsymbol{\mu}_{y=0} = \begin{bmatrix}-1\\0\end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix}1 & -1\\1 & 2\end{bmatrix}, \quad \boldsymbol{\mu}_{y=1} = \begin{bmatrix}1\\0\end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix}1 & -1\\1 & 2\end{bmatrix}$

D. $\boldsymbol{\mu}_{y=0} = \begin{bmatrix}-1\\0\end{bmatrix}, \boldsymbol{\Sigma}_{y=0} = \begin{bmatrix}1 & -1\\1 & 2\end{bmatrix}, \quad \boldsymbol{\mu}_{y=1} = \begin{bmatrix}1\\0\end{bmatrix}, \boldsymbol{\Sigma}_{y=1} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$

Class-conditional Gaussian Distributions

Iris dataset:

- Class distribution: $Y \sim Bern(\phi)$ (or Categorical)
- Class conditional distribution: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$

- Naïve Bayes assumption: Σ are all diagonal
- Linear Decision Boundary: $\Sigma_{Y=0} = \Sigma_{Y=1}$
- Quadradic Decision Boundary: $\Sigma_{Y=0} \neq \Sigma_{Y=1}$

Generative + MAP

MLE vs MAP vs Generative vs Discriminative

	Maximum likelihood estimation	Maximum *a posteriori* estimation
Descriminative Models only the conditional likelihood, $\prod p(y \mid x, \theta)$	$p(\mathcal{D}^* \mid \theta)$ * conditional likelihood $= \prod_{i=1}^{N} p(y^{(i)} \mid x^{(i)}, \theta)$	$p(\theta)\, p(\mathcal{D}^* \mid \theta)$ * conditional likelihood $= p(\theta) \prod_{i=1}^{N} p(y^{(i)} \mid x^{(i)}, \theta)$
Generative Models the full joint likelihood, $\prod p(x, y \mid \theta)$	$p(\mathcal{D} \mid \theta)$ * actual likelihood $= \prod_{i=1}^{N} p(x^{(i)}, y^{(i)} \mid \theta)$ $= \prod_{i=1}^{N} \underbrace{p(x^{(i)} \mid y^{(i)}, \theta)}_{\text{class-conditional}} \underbrace{p(y^{(i)} \mid \theta)}_{\text{class prior}}$	$p(\theta)\, p(\mathcal{D} \mid \theta)$ * actual likelihood $= p(\theta) \prod_{i=1}^{N} p(x^{(i)}, y^{(i)} \mid \theta)$ $= p(\theta) \prod_{i=1}^{N} \underbrace{p(x^{(i)} \mid y^{(i)}, \theta)}_{\text{class-conditional}} \underbrace{p(y^{(i)} \mid \theta)}_{\text{class prior}}$

Reminder: Prior Distributions for MAP

If the prior $p(\theta)$ is uniform, then MLE and MAP are the same!

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p\left(y^{(i)} \mid \phi\right) p(\phi) \;=\; \phi^{N_{y=1}} (1-\phi)^{N_{y=0}}\, p(\phi)$$

Conjugate priors: when the prior and the posterior distributions are in the same family

Bernoulli likelihood with a Beta prior has Beta posterior

Categorical likelihood with a Dirichlet prior has Dirichlet posterior

Gaussian likelihood with a Gaussian prior has Gaussian posterior

https://www.desmos.com/calculator/kr7m2m6cf7