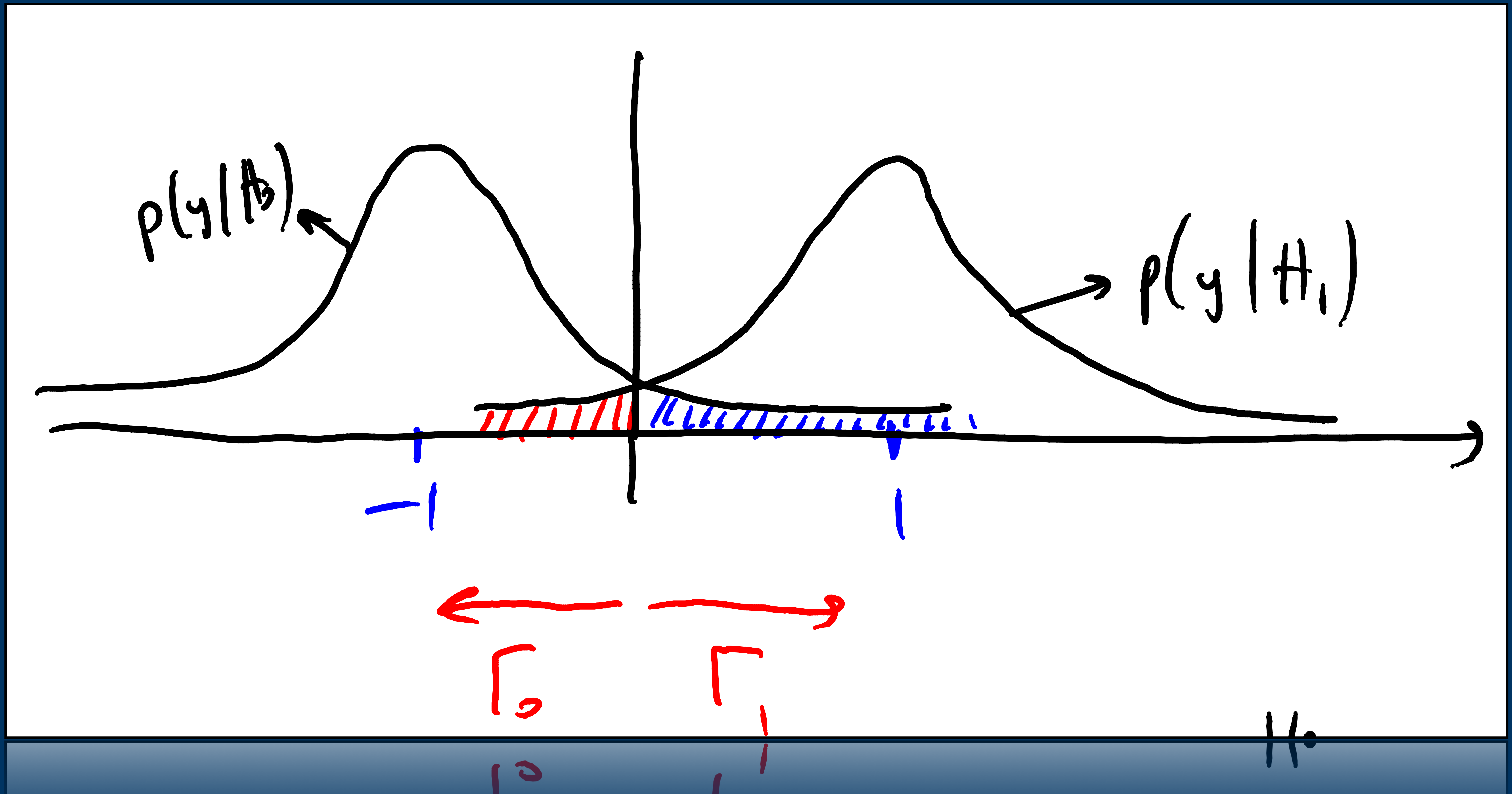


Naive Bayes Classification

CSCI-P556 Applied Machine Learning
Lecture 10

D.S. Williamson

Agenda and Learning Outcomes

Today’s Topics

- **Topics**:
  - Finish Probability Review
  - Naive Bayes Classification
- **Announcements**
  - Create private repos. **~9 of you haven’t**. See Piazza
  - Commit to Github early and often
  - Put name of partners as a text comment. Answer to questions as text comments as well
  - Each non-submitting member commit a text file (or something similar) that specifies whose implementation to grade. ***Only one submission per group will be graded***.

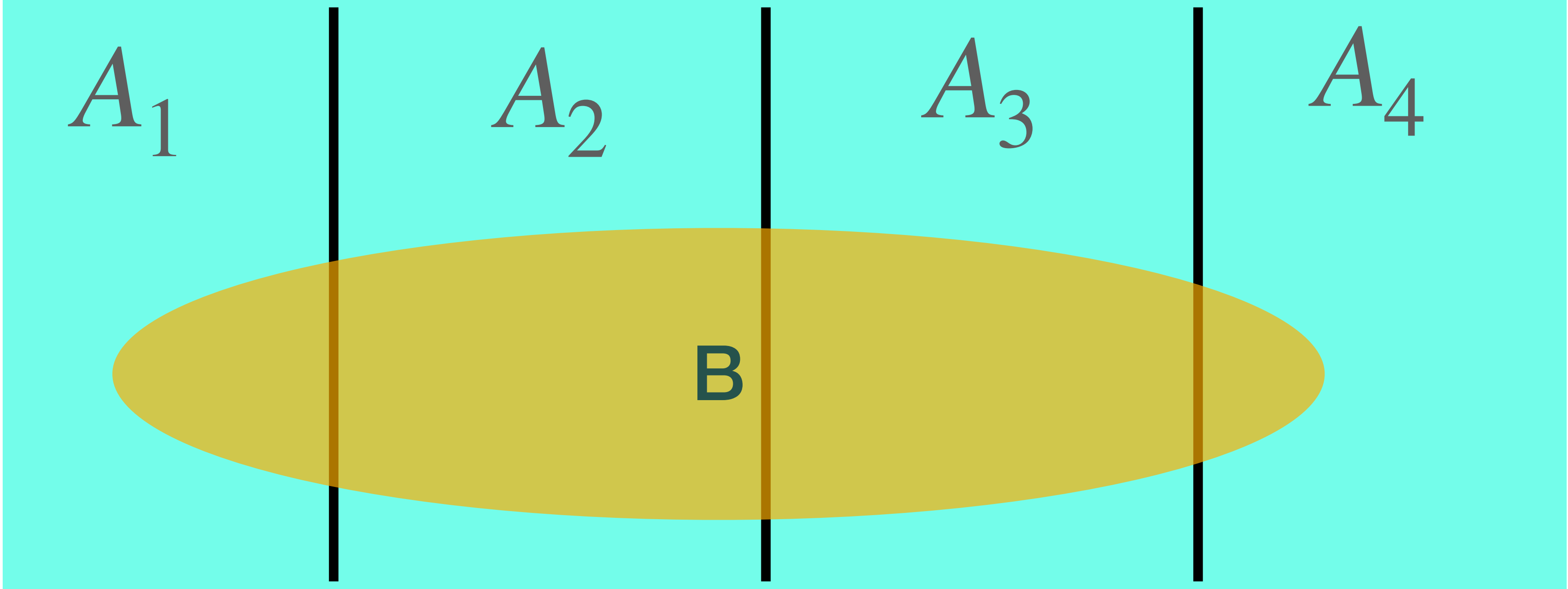
$$P(A_i \mid B) = \frac{P(A_i, B)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{P(B)}$$

Bayes Rule (Theorem)

Conditional Probability can be re-written

- A_1, A_2, … form a partition and represent distinct sets of possible outcomes
  - $P(A_j) > 0$ for all j
- Let B be any event s.t. $P(B) > 0$

All possible outcomes



$P(A_i,B)$ -> Law of Multiplication

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{\infty} P(A_j)P(B|A_j)}$$

for any i=1,2,…

P(B) -> Law of Total Probability

Bayes Rule: Example

Suppose

- 50% of voters are Democrats
- 30% of voters are Republican
- 20% of voters are Independent

- 40% of Democrats voted for candidate X
- 80% of Republicans voted for candidate X
- 50% of Independents voted for candidate X

What fraction of candidate X's votes came from Republications?

- Solution
  - A_1 = {Democrats}, A_2 = {Republican}, A_3 = {Independent}
  - B = { voted for X}
  - $P(A_2 | B)$ = ?
  - $P(A_1) = 0.5$, $P(A_2) = 0.3$, $P(A_3) = 0.2$
  - $P(B | A_1) = 0.4$, $P(B | A_2) = 0.8$, $P(B | A_3) = 0.5$
  - A_1, A_2, and A_3 form a partition

=>

$$\begin{aligned} P(A_2|B) &= \frac{P(A_2, B)}{P(B)} \\ &= \frac{P(A_2)P(B|A_2)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)} \\ &= \frac{0.3 * 0.8}{0.5 * 0.4 + 0.3 * 0.8 + 0.2 * 0.5} = \frac{4}{9} \end{aligned}$$

Independence

Independence

- X and Y are independent if and only if:

$$f(x, y) = f_X(x) f_Y(y) \quad \text{for CRVs}$$
$$P(x, y) = P_X(x) P_Y(y) \quad \text{for DRVs}$$

- Equivalent condition
  - if X and Y are independent for all x and y, then

$$f(y|x) = \frac{f(x, y)}{f(x)} = \frac{f(x)f(y)}{f(x)} = f(y)$$

- Similarly
  - X and Y are independent ⇔ *f(x,y) = g(x)h(y)* for some functions *g* and *h*
  - In other words,
    - if the joint distribution (PMF or PDF) can be written as a product of a function of RV *X* and a function of RV *Y*, then *X* and *Y* are independent
    - If X and Y are independent, then the joint distribution is a product of a function of *X* and a function of *Y*

Independence: Example

- Are X and Y independent?

$$f_{XY}(x,y) = \begin{cases} 2e^{-(x+2y)}, & \text{if} \quad 0 \le x, 0 \le y \\ 0, & \text{otherwise} \end{cases}$$

  - Solution
    - Can show that

$$\begin{aligned} 2e^{-(x+2y)} &= e^{-x}(2e^{-2y}) \\ &= f_X(x) f_Y(y) \end{aligned}$$

  - X and Y are independent

Conditional Independence

- Have random variables *X, Y, Z*

- Consider *P(x,y | z)*

- *X* and *Y* are <u>conditionally independent</u> given *Z* if:

$$P(x,y|z) = P(x|z)P(y|z)$$

Covariance

- <u>Definition</u>
  - The covariance between two RVs (*X* and *Y*) is a measure of their association

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] \quad \text{where } \mu_x = E[X]$$
$$\mu_y = E[Y]$$

- So

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_x)(Y-\mu_y)] \quad \text{where} \\ &= E[XY - \mu_x Y - \mu_y X + \mu_x\mu_y] \\ &= E[XY] - \mu_x E[Y] - \mu_y E[X] + \mu_x\mu_y \\ &= E[XY] - 2\mu_x\mu_y + \mu_x\mu_y \\ &= E[XY] - \mu_x\mu_y \end{aligned}$$

Sign of Cov(X,Y)

- <u>If Cov(X,Y) > 0</u>
  - High values of X tend to occur with high values of Y
  - Low values of X tend to occur with low values of Y

$$Cov(X, Y) = E[XY] - \mu_x \mu_y$$

- <u>If Cov(X,Y) < 0</u>
  - High values of X tend to occur with low values of Y
  - Low values of X tend to occur with high values of Y
- <u>If Cov(X,Y) = 0</u>
  - X and Y are <u>uncorrelated</u>

Covariance and Correlation

- Correlation of X and Y

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \qquad -1 \leq \rho_{XY} \leq 1$$

- <u>Useful facts</u>
  - $\rho_{XY} = 1$ <=> perfect positive linear association.
    with probability of 1, Y = aX + b for some $a > 0$
  - $\rho_{XY} = -1$ <=> perfect negative linear association.
    with probability of 1, Y = aX + b for some $a < 0$

Covariance and Correlation

- Correlation of X and Y

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

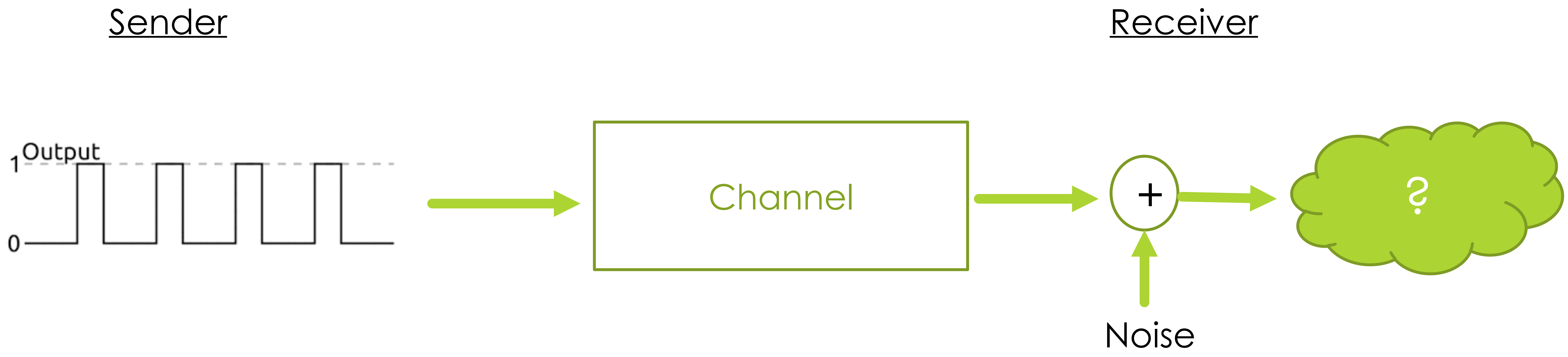
- <u>Useful facts</u>
    - If X and Y are independent, then Cov(X,Y) = ρ_{XY} = 0
        - <u>Proof</u>:
        - Cov(X,Y) = E[XY] – E[X]E[Y]
        - = E[X]E[Y] – E[X]E[Y] = 0
        - The reverse of this is not true: Cov(X,Y) = 0 then X and Y are independent (wrong!)

Naive Bayes Classification

A Classification Example

Digital Communication

- A digital pulse train is transmitted over some channel



- Let:
  - y(t) be the transmitted pulse (label)
  - x(t) be the received signal (feature/observation)

<u>Suppose x(t) is observed as:</u>

- $x(t) = y(t) + N(t)$, if '0' is sent
- $x(t) = N(t) - y(t)$, if '1' is sent

$0 \leq t \leq T$

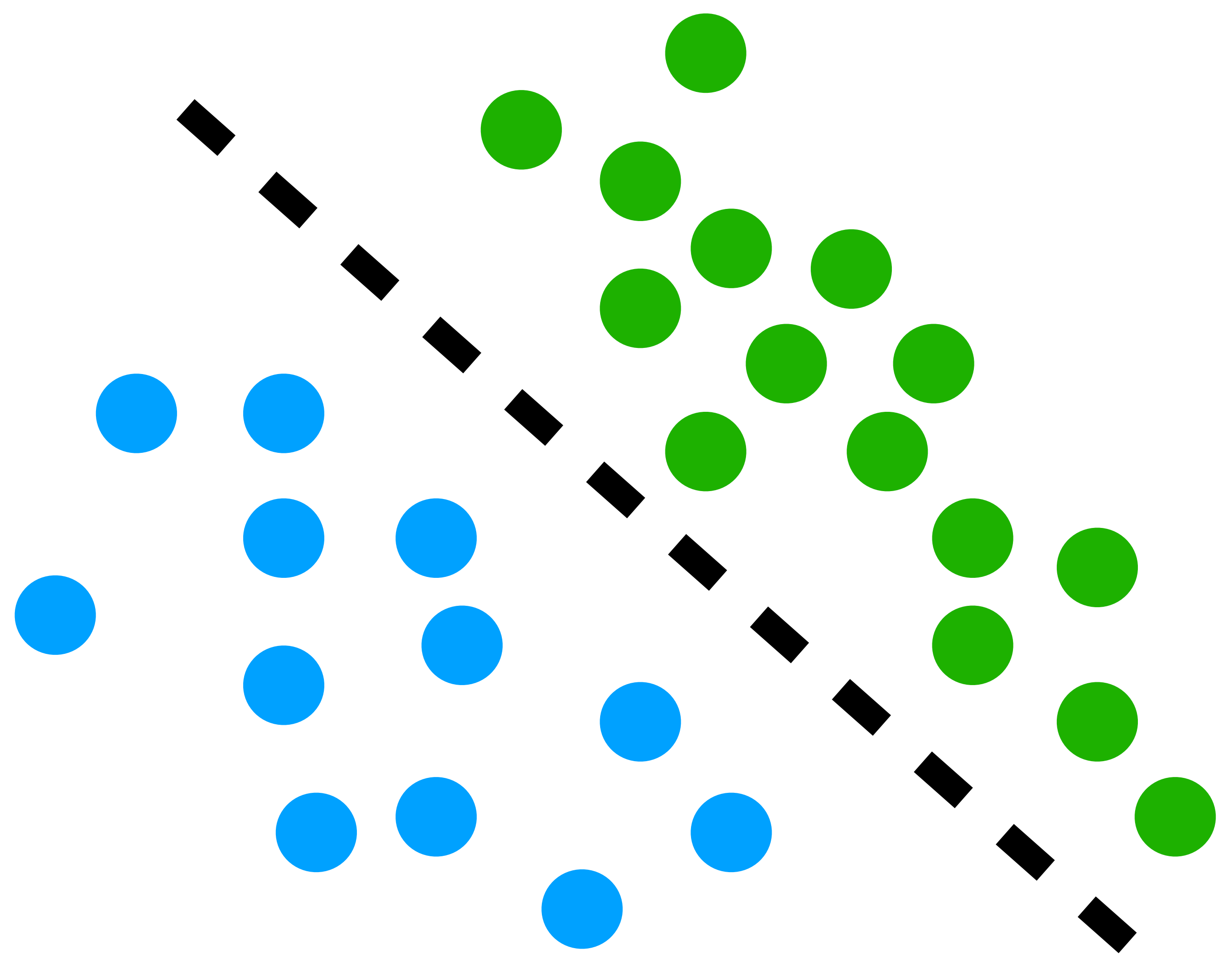
Must decide, from the observation, whether a "0" or "1" bit is transmitted

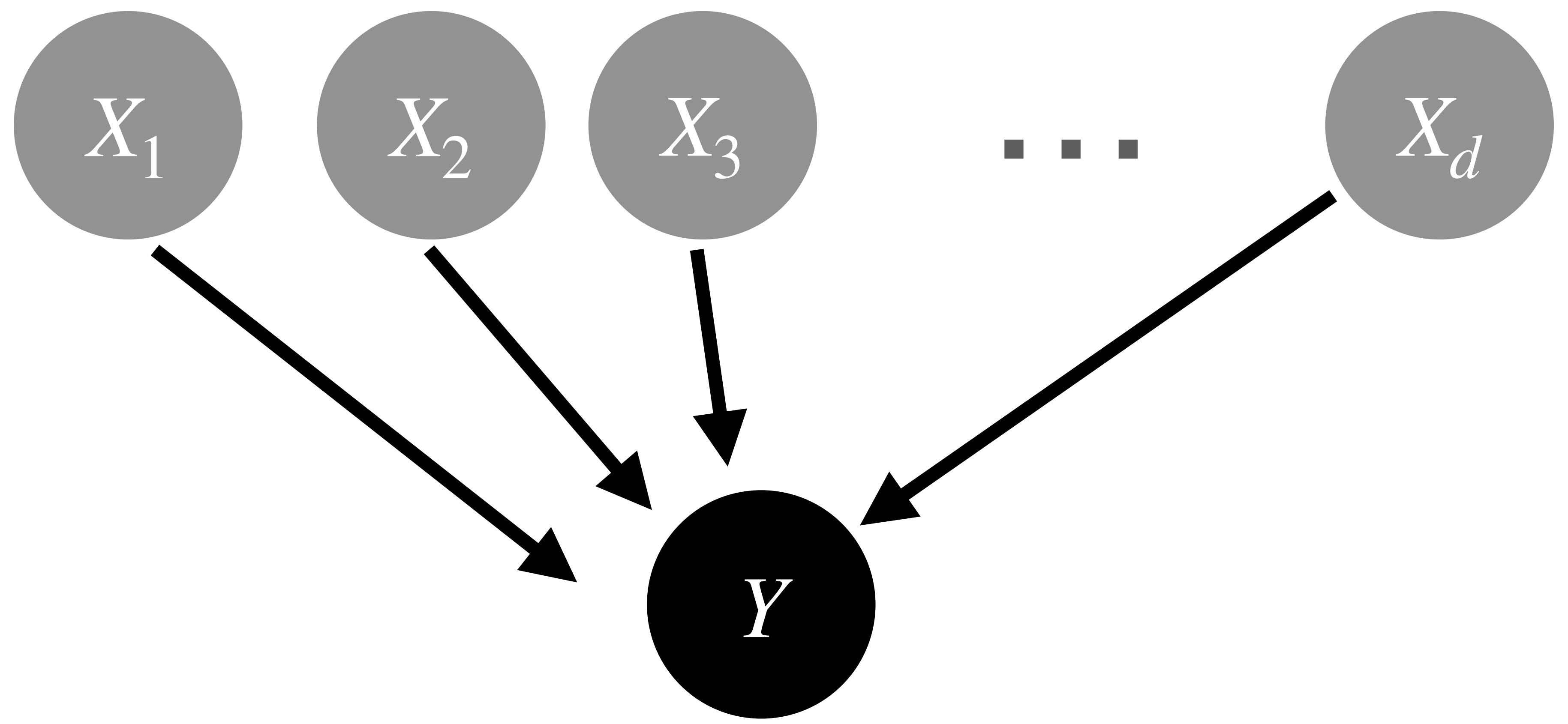
Generative vs <u>Discriminative</u> Classification

Two different approaches to performing classification

- Two probabilistic approaches for classification that use the (discrete) input features, ***x***, and corresponding labels (or classes), *y*, differently

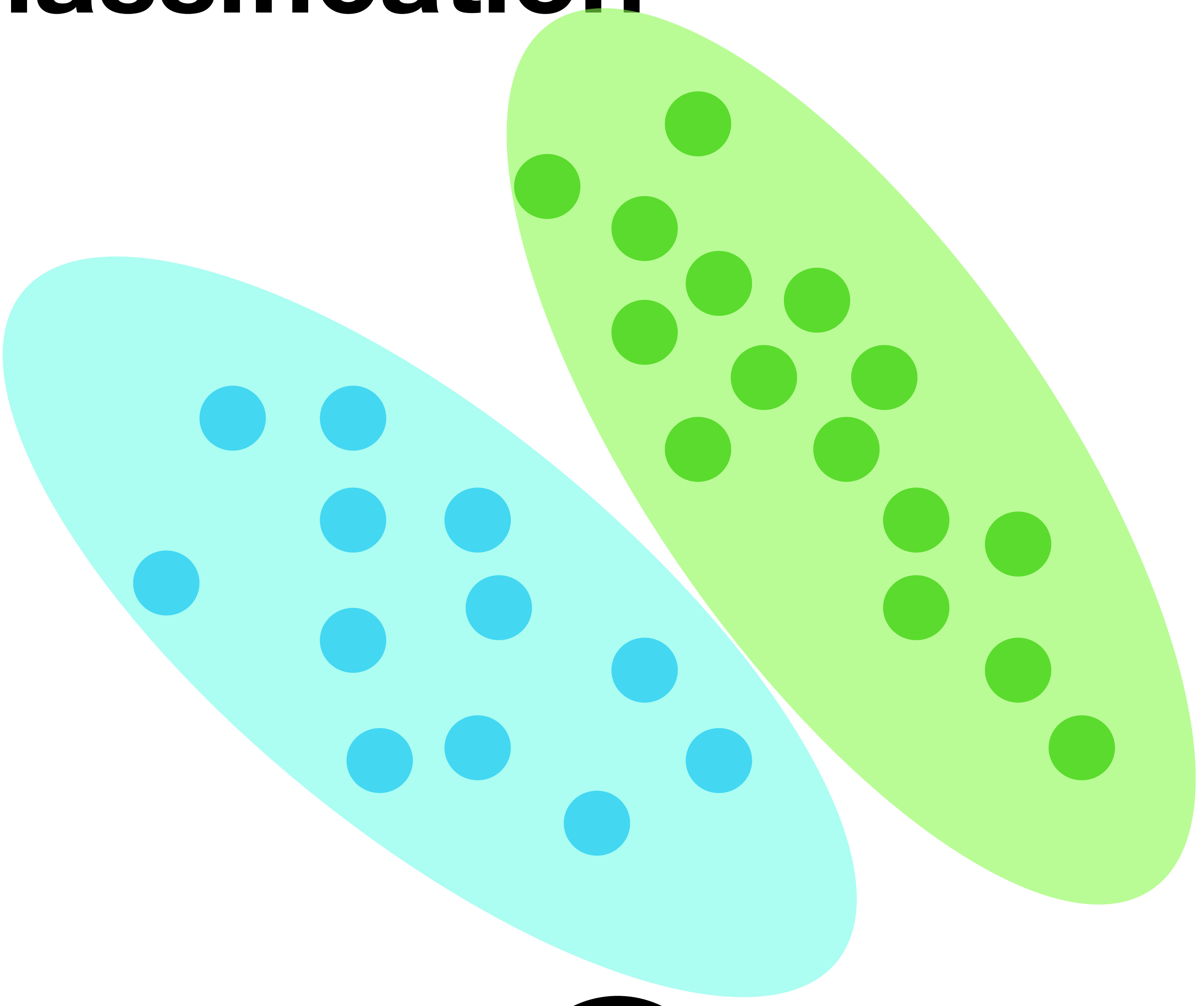
- ***<u>Discriminative methods</u>*** model P(y|**x**) directly (e.g. Logistic Regression), by focusing on the task of categorizing data
  - **Goal**: Given the feature, what is the probability of observing this class?
  - Captures <u>differences</u> between categories. Eg, what differentiates birds and mammals?
  - Assumes value of label (or class) depends on the features
  - Typically more efficient and simpler





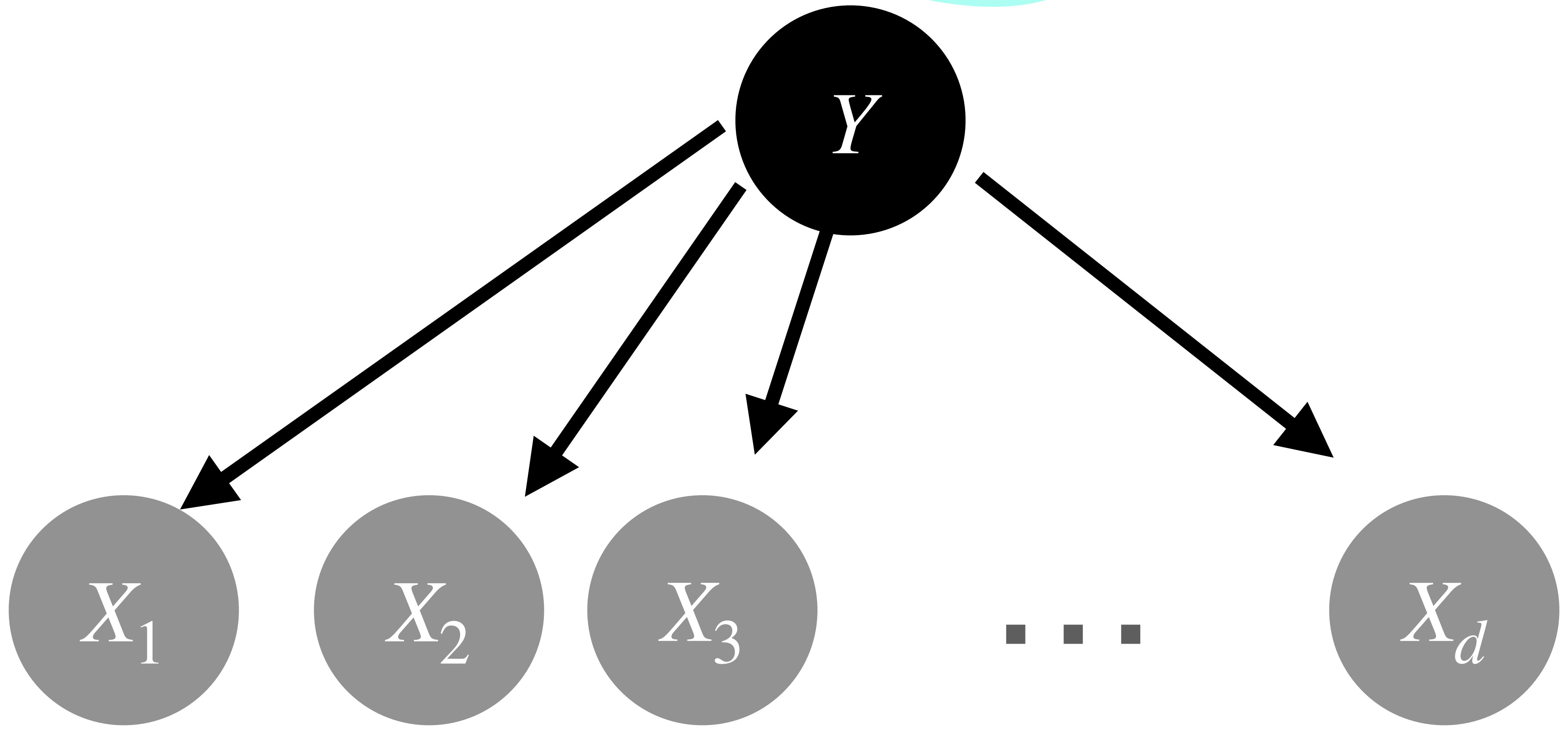
Generative vs Discriminative Classification

Two different approaches to performing classification



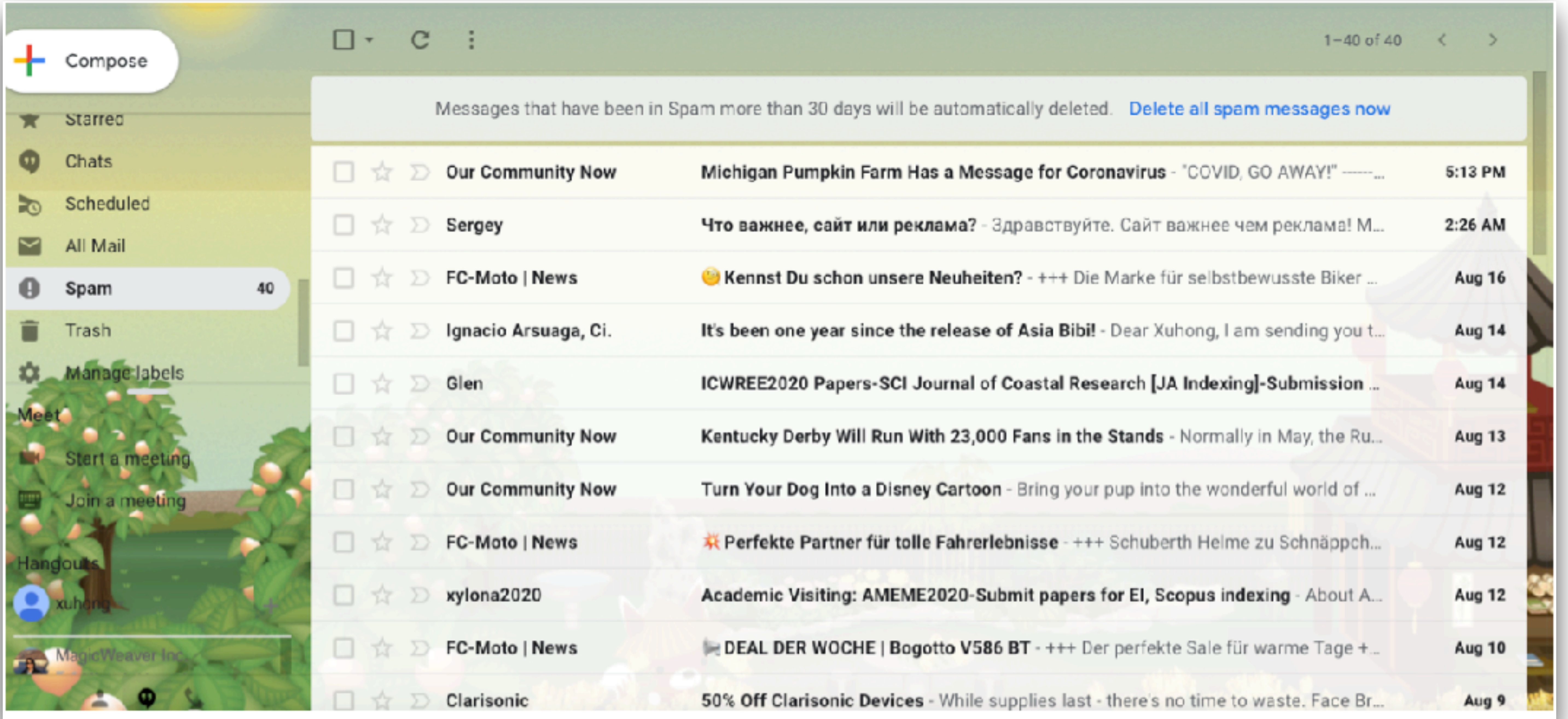
- Two probabilistic approaches for classification that use the (discrete) input features, ***x*,** and corresponding labels (or classes), *y*, differently

- ***Generative methods*** model P(**x**|y) and P(y)
  - **Goal**: Given the label, what is the probability of this feature? What is the probability of the label?
  - Assumes input feature can be "created" if the label is known
  - Describes probability distributions for all features
  - Stochastically create a plausible feature vector
  - Make a model that generates positives (e.g. Class A). Make a model that generates negatives (Class B).
  - **Classify** a test example based on which is more likely to generate it (e.g. Naïve Bayes)



Success Story of Naïve Bayes: Spam Filter

- The naive Bayes classifier is widely used for text data
- ***Problem***:
  - We want to classify email messages into spam and non-spam categories
  - Our training set is a set of emails that has been classified manually into the two categories
- ***First questions***: How do we represent an email using a feature vector ***x*** - what features should we use?



A Bayes Classifier

Example: Spam Email Detection

- Our **features** are a binary encoding of possible words that are in the email.
  - Our vocabulary has n words. Hence, there are 2^n possible values for x
  - Feature is n-dimensional, where value at i-th index of feature, x_i, is 1 if word is in the email and 0 otherwise.
- **Goal**: Learn a Bayes Classifier, hence we need to model:
  - P(**x**|y=0) (e.g. probability of feature given not a spam email)
  - P(**x**|y=1) (e.g. probability of feature given is a spam email)
  - P(y=0) (e.g. probability of not having spam email)
  - P(y=1) (e.g. probability of spam email)

A Bayes Classifier

Example: Spam Email Detection

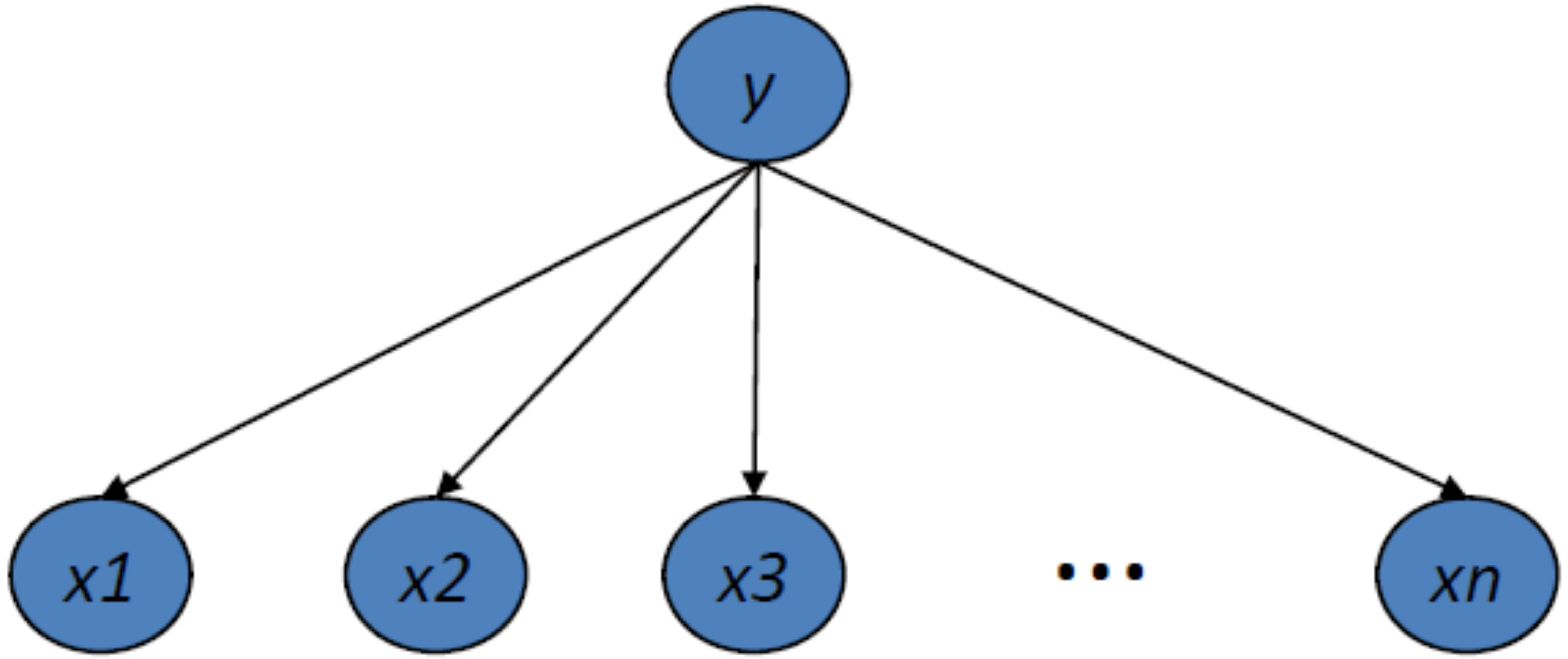
- **Problem**: Modeling P(**x**|y) (for y = 0 and 1) requires too many parameters (e.g. $2^*(2^n - 1)$). It's a multinomial distribution.
- **Solution**: Apply the Naive Bayes assumption, which assumes that x_i's are conditionally independent (e.g. assume words are conditionally independent)

$$P(\mathbf{x} \mid y) = P(x_1, x_2, \cdots, x_n \mid y) = \prod_{i=1}^{n} P(x_i \mid y)$$

The number of parameters for P(**x**|y) is now 2*n. Why?

Hence *avoids* estimating joint probabilities across features (eg., $x_1 \wedge x_2$)

Naïve Bayes



- A ***generative model*** — an email is generated as follows:
  - Determine if it is a spam email or not according to P(y) (Bernoulli)
  - Determine if each word x_i in the vocabulary is contained in the message *independently* according to $P(x_i \mid y)$ (Bernoulli)
- For this model, ***<u>we need to learn</u>***:
  - <u>For y</u>: $P(y=1)$ (and $P(y=0)$). These are our ***<u>prior</u>*** probabilities
  - <u>For x_i</u>: $P(x_i = 1 \mid y = 1)$ and $P(x_i = 1 \mid y = 0)$. These are are ***<u>class conditional probabilities</u>*** for i = 1,…, n.
    - Can then compute $P(x_i = 0 \mid y = 1)$ and $P(x_i = 0 \mid y = 0)$. How?

Maximum Likelihood Estimate (MLE) for Naïve Bayes

- Suppose our training set contained N emails, the ***maximum likelihood estimate*** of the parameters are:

$$P(y=1) = \frac{N_1}{N}, \text{ where } N_1 \text{ is the number of spam emails}$$

$$P(x_i = 1 \mid y = 1) = \frac{N_{i|1}}{N_1},$$

i.e., the fraction of spam emails where x_i appeared

$$P(x_i = 1 \mid y = 0) = \frac{N_{i|0}}{N_0}$$

i.e., the fraction of the nonspam emails where x_i appeared

What if x_i is Multinomial?

- If x_i is discrete with more than two possible values $\{v_1, \cdots, v_m\}$, $P(x_i \mid y)$ can be described by a conditional probability table
- Really only needs m-1 rows since rows sum to 1
- More columns can be added for the multi-class case
- $P(x_i = v_j \mid y = k) = \dfrac{N_{ij|k}}{N_k}$, i.e., the fraction of class k examples where x_i took value v_j

	y = 0	y = 1
$x_i = v_1$	$P(x_i = v_1 \mid y = 0)$	$P(x_i = v_1 \mid y = 1)$
$x_i = v_2$	$P(x_i = v_2 \mid y = 0)$	$P(x_i = v_2 \mid y = 1)$
…	…	…
$x_i = v_3$	$P(x_i = v_m \mid y = 0)$	$P(x_i = v_m \mid y = 1)$

Learning and Classification

- ***Learning***: Need to estimate the following probability distributions (via counting from data)
  - Prior distribution of y: $P(y)$
  - Class conditional distribution of x_i: $P(x_i \mid y)$
- ***Classification/Predicting***:
  - Given $x = (x_1, x_2, \cdots, x_d)$, compute $P(y \mid \mathsf{x})$ for y = 0 and y = 1.
  - Apply decision theory to make final prediction of y

$$P(y \mid \mathsf{x}) = \frac{P(y)P(\mathsf{x} \mid y)}{P(\mathsf{x})} \propto P(y) \prod_i P(x_i \mid y)$$

Naive Bayes Classification

Choose the class that is optimal

- Once the likelihood and prior probabilities are determined, Naive Bayes performs classification by choosing the class with the larger weighted (by the prior) likelihood

$$P(y=0)P(\mathbf{x}\,|\,y=0) \lesseqgtr_0^1 P(y=1)P(\mathbf{x}\,|\,y=1)$$

- This can be reformulated as below, where this is known as the ***likelihood ratio test (LRT)***

$$\frac{P(\mathbf{x}\,|\,y=0)}{P(\mathbf{x}\,|\,y=1)} \lesseqgtr_0^1 \frac{P(y=1)}{P(y=0)}$$

- If the prior probabilities are equal, then this becomes: $P(\mathbf{x}\,|\,y=0) \lesseqgtr_0^1 P(\mathbf{x}\,|\,y=1)$, which is known as the ***Maximum Likelihood Decision Rule***

Naive Bayes Classification

Practical Considerations

- Recall that the likelihood of $\mathbf{x}$ given y (e.g. $P(\mathbf{x} \mid y)$) for our email example assumes independence for the words

$$P(\mathbf{x} \mid y = 0) = \prod_i P(x_i \mid y = 0) = P(x_1 \mid y = 0) \times P(x_2 \mid y = 0) \times \cdots \times P(x_n \mid y = 0)$$

- In practice, taking the product of probabilities can lead to ***underflow***, where the number is too small to be represented in the computer. To avoid this, we often operate in the log-domain so that the log-probabilities are added instead

$$\log\left(P(\mathbf{x} \mid y = 0)\right) = \log\left(P(x_1 \mid y = 0)\right) + \log\left(P(x_2 \mid y = 0)\right) + \cdots + \log\left(P(x_n \mid y = 0)\right)$$

- The ratio test then becomes: $$\frac{\log\left(P(x_1 \mid y = 0)\right) + \cdots + \log\left(P(x_n \mid y = 0)\right)}{\log\left(P(x_1 \mid y = 1)\right) + \cdots + \log\left(P(x_n \mid y = 1)\right)} \lessgtr_0^1 \frac{P(y = 1)}{P(y = 0)}$$

Problem with (MLE) Naive Bayes Classification

- Many words are rare, resulting in poor probability estimates for certain words

- Consider the spam example:
  - Suppose in our training set “Mahalanobis” appears in a non-spam e-mail and never appears in a spam e-mail
  - Suppose also that “XXX” appears in a spam message but not a non-spam messages
  - Now suppose we get a new message **x** that contains both words
- We will have that $P(\mathbf{x} \mid y) = \prod_i P(x_i \mid y) = 0$ for both y =0 and y = 1
  - Because P(“Mahalanobis” | y = 1) = 0 and P(“XXX” | y = 0) = 0
- Given this limited training data, Naive Bayes (via MLE) can result in probabilities of 0 or 1. Such extreme probabilities are “too strong” and cause problems.
  - Use Smoothing techniques to help correct this (e.g. Laplace Smoothing)
  - More on this during next homework.

Example: Digital Communication

A digital pulse train is transmitted over some channel

- Decide `0’ or `1’ based on the observation x.
- Under:
  - y=0, x ~ $N(-1,1)$ (e.g. $f(x \mid y=0)$ follows a Gaussian distribution with mean -1 and variance of 1.
  - y=1, x ~ $N(1,1)$ (e.g. $f(x \mid y=1)$)
- Let $P(y=0) = P(y=1) = 0.5$

Find the classifier?

Example: Digital Communication

A digital pulse train is transmitted over some channel

- Decide `0' or `1' based on the observation x.
- For:
  - *y=0, x ~ N(-1,1)*. Hence, f(x | y=0) follows a Gaussian distribution with mean -1 and variance of 1.
  - y=1, x ~ N(1,1) (e.g. f(x | y =1)
- Let P(y=0) = P(y=1) = 0.5

Find the classifier?

$$\frac{f(x\,|\,y=0)}{f(x\,|\,y=1)} \lessgtr_0^1 \frac{P(y=1)}{P(y=0)} \quad \Rightarrow \quad \frac{f(x\,|\,y=0)}{f(x\,|\,y=1)} \lessgtr_0^1 1$$

$$\frac{f(x\,|\,y=0)}{f(x\,|\,y=1)} = \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x+1)^2}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-1)^2}} = e^{-2x}$$

$$\Rightarrow \quad e^{-2x} \lessgtr_0^1 1$$

$$\Rightarrow \quad \log\left(e^{-2x}\right) \lessgtr_0^1 \log(1)$$

$$-2x \lessgtr_0^1 0 \quad \Rightarrow \quad 0 \lessgtr_0^1 x$$

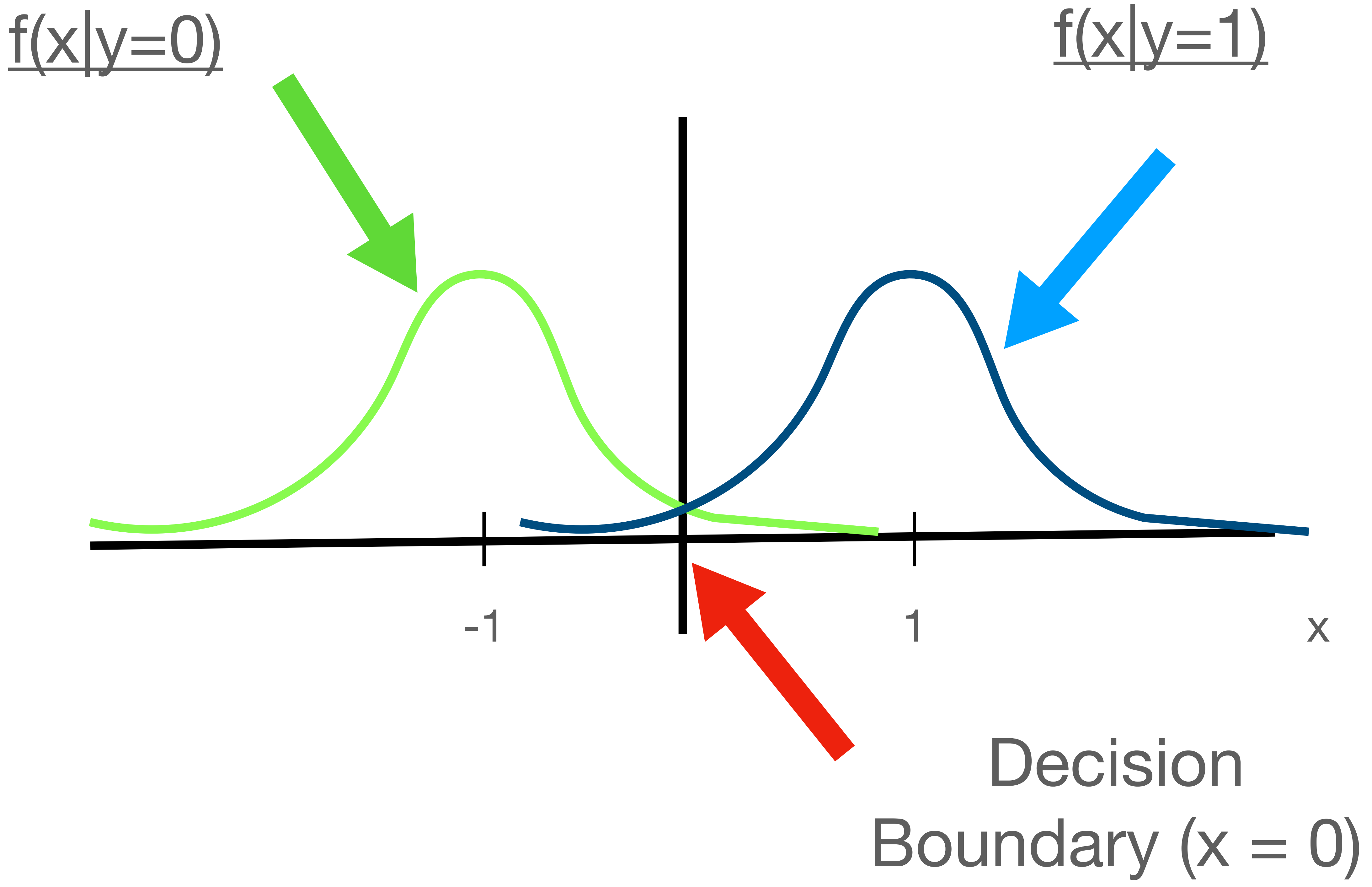
Example: Digital Communication

A digital pulse train is transmitted over some channel

- Decide `0' or `1' based on the observation x.
- For:
  - *y=0, x ~ N(-1,1)*. Hence, f(x | y=0) follows a Gaussian distribution with mean -1 and variance of 1.
  - y=1, x ~ N(1,1) (e.g. f(x | y =1)
- Let P(y=0) = P(y=1) = 0.5

Find the classifier?

$$0 \underset{>_0}{\overset{\leq_1}{}} x$$



What about the multi-class case?

Instead of having two classes (0 and 1), now have multiple class options (0 to K)

- **How is Naive Bayes multi-class classification performed?**
- After simplifying the problem, the classifier becomes

$$\hat{y} = \arg\max_k f(x|y=k)P(y=k)$$

In other words, you choose the class with the maximum posterior (or likelihood)

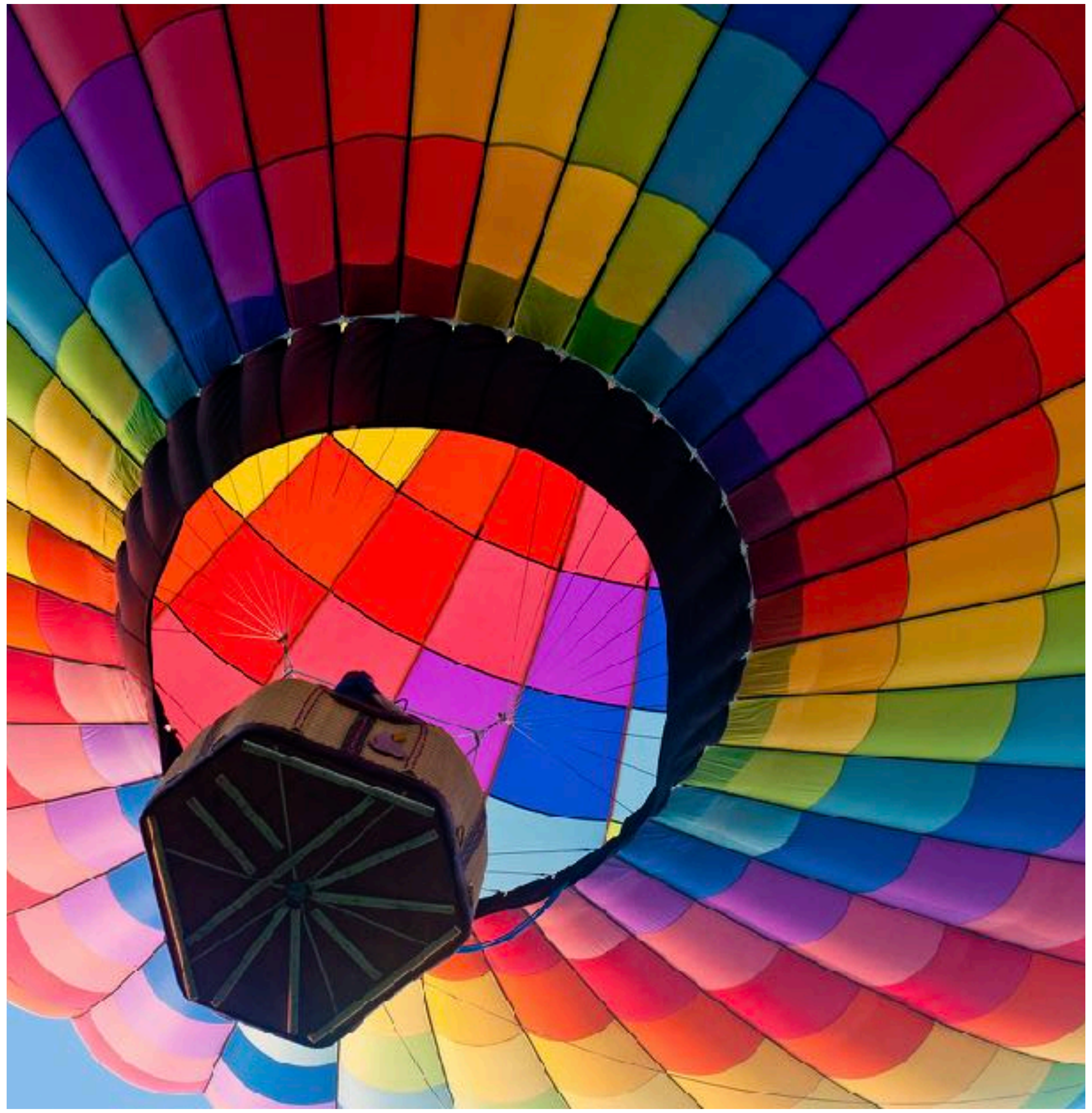
Summary - Naive Bayes Classification

- Generative classifier
  - Learn P(x|y) and P(y).
  - Use Bayes rule to compute P(y|x) for classification

- Assumes conditional independence between features given labels
  - Greatly reduces the number of parameters
  - Referred to as the ***Naive assumption***

- Batch learning, but can be turned into online learning
  - Incrementally update the various probability estimates

- Often a good “first thing” to try. It often works surprisingly well.



Next Class: Linear Regression