2^D 50 features

Naïve Bayesian Inference/Classification

Bayes Rule

- If you recall from probability, Bayes' provided another way to write the posterior $P(y = i|f = x)$:

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

- In Bayes' Rule we call
  - $P(y = i|f = x)$ the posterior (what we want)
  - $P(y = i)$ the prior (probability that $y = i$)
  - $P(f = x|y = i)$ the likelihood (likelihood of generating x given y)
  - $P(f = x)$ the evidence

Bayes Rule

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

- The prior is easy to come by if we have data
  - The prior $P(y = i)$ is just the percentage of samples that have their class $y = i$
- How about the likelihood $P(f = x|y = i)$?
- Obviously we can get it if we have a robust joint distribution table: $P(f = x|y = i) = \frac{P(f=x,y=i)}{P(y=i)}$
- But what if we don't have enough data for a robust joint distribution table?

Naïve Bayes Probability

- If we make the assumption that the features are *conditionally independent* (that is, that given one feature conditioned on our class we can't say anything about the other features), then we can re-write $P(f = x|y = i)$ as:

$$P(f = x|y = i) \approx \prod_{k=1}^{D} P(f_k = x_k|y = i)$$

- This is a large assumption, but one that is often made.
- So we can now approximate our posterior as:

$$P(y = i|f = x) = \frac{P(y = i)P(x|y = i)}{P(f = x)} \approx P(y = i)\prod_{k=1}^{D} P(f_k = x_k|y = i)$$

- We call this approach, *Naïve Bayes*

Naïve Bayes Probability

$$P(y = i|f = x) \approx P(y = i) \prod_{k=1}^{D} P(f_k = x_k|y = i)$$

- Note that this formula doesn't have the evidence $P(x)$ in it!
  - That's because since we're approximating the numerator based on the conditional independent assumption, the overall Bayes rule doesn't hold.
- Fortunately since we know that since the sum of the posteriors over all classes should be one, if we need an actual probability we can again just divide by the sum of this equation over all classes:

$$P(y = i|f = x) \approx \frac{P(y = i) \prod_{k=1}^{D} P(f_k = x_k|y = i)}{\sum_j P(y = j) \prod_{k=1}^{D} P(f_k = x_k|y = j)}$$

Example

- Let's try to determine if an object is a banana, an orange, or something else based on its length, sweetness, and color.
  - Where length, sweetness and color are all binary features (isLong?, isSweet?, isYellow?)
- Below is a table showing observations of 1000 pieces of fruit
  - For instance, there are 500 Bananas, and 400/500 of them are considered "Long"

Fruit	Long	Sweet	Yellow	Total
Banana	400	350	450	500
Orange	0	150	300	300
Other	100	150	50	200
Total	**500**	**650**	**800**	**1000**

Example

- How about with Naïve Bayes?

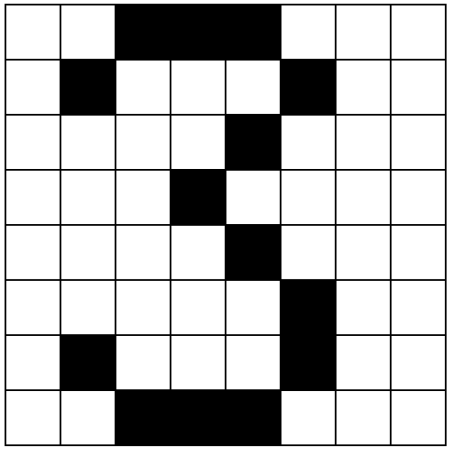
$$P(B|L, \neg S, Y) \approx P(B)P(L|B)P(\neg S|B)P(Y|B)$$

- But if I want an actual probability we also need to compute this for the Orange and Other class than divide by their sum…

Fruit	Long	Sweet	Yellow	Total
Banana	400	350	450	500
Orange	0	150	300	300
Other	100	150	50	200
Total	**500**	**650**	**800**	**1000**

Example – Digit classification

- Input: pixel grids

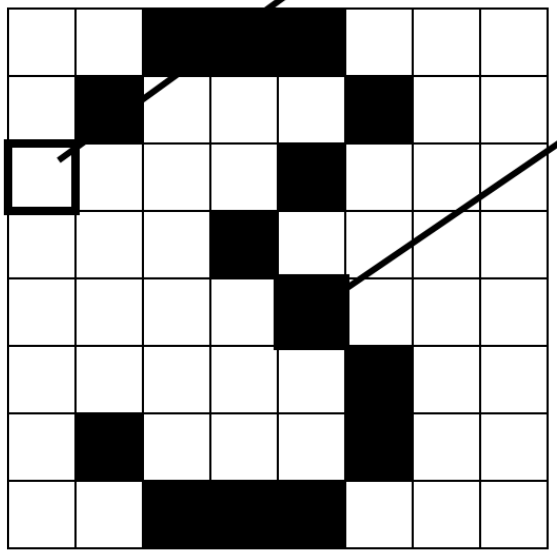


- Output: a digit 0-9

Digit Classification

$P(Y)$

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



$P(F_{3,1} = on|Y)$

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

$P(F_{5,5} = on|Y)$

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

Careful…

$P(\text{features}, C = 2)$

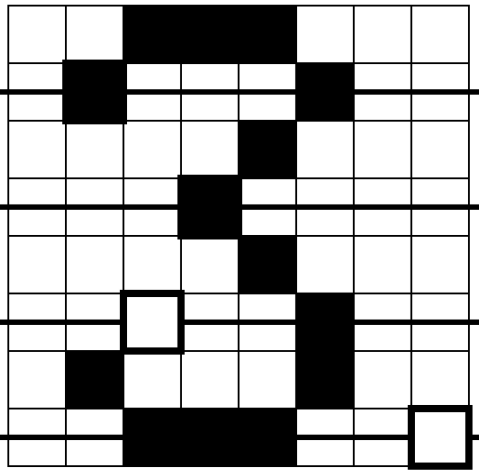
$P(C = 2) = 0.1$

$P(\text{on}|C = 2) = 0.8$

$P(\text{on}|C = 2) = 0.1$

$P(\text{off}|C = 2) = 0.1$

$P(\text{on}|C = 2) = 0.01$



$P(\text{features}, C = 3)$

$P(C = 3) = 0.1$

$P(\text{on}|C = 3) = 0.8$

$P(\text{on}|C = 3) = 0.9$

$P(\text{off}|C = 3) = 0.7$

$P(\text{on}|C = 3) = 0.0$

2 wins!!

Inference vs Naïve Bayes

- Which should we use?
- Depends on what you have
  - Ideally use regular inference
  - If we have even less joint information and we can make an independence assumption then we can try naïve inference

Continuous Valued Data

Categorical vs Continuous Valued Data

- Each feature can typically fall into one of two categories:
  1. Categorical– The feature can have one of M possible values (categories, enumerations, finite possible values)
  2. Continuous – The features can have any value!
- In all our inference work we had to essentially count how many times something occurred in order to compute its probability
  - This is only feasible for categorical data
- What if our data is continuous?

What if we have continuous x_k?

- The general form of Naïve Bayes is:

$$P(y = i | f = x) \propto P(y = i) \prod_{j=1}^{D} P(f_k = x_k | y = i)$$

- If a feature is enumerated, then extract the individual probabilities $P(f_k = x_k | y = i)$ from the observed data.
- Alternatively maybe we can decide on an equation for $P(f_k | y = i)$!

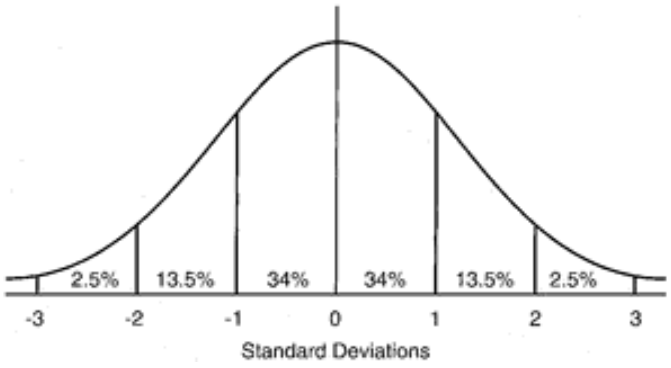
Gaussian Distributions

- If we decide that the feature is normally distributed (i.e a Gaussian, bell-shaped curve, etc..) then we can get something proportional to the probability $P(f_k = x_k | y = i)$ as:

$$P(f_k = x_k | y = i) \propto \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x_k - \mu_i)^2}{2\sigma_i^2}}$$

  - Where μ_i is the expected or mean value of feature f_k for data from class $y = i$
  - And σ_i is the standard deviation of the feature f_k for data from class $y = i$

- Note: This equation is referred to as the in probability as the probability density function (pdf)



Gaussian Naïve Bayes: Continuous x, Discrete y

- How do we train a Naïve Bayes classifier with continuous features?
- For each discrete class C_i
  - First estimate the prior $P(y = i)$
  - For each attribute k, *estimate* $P(f_k = x_k | y = i)$ as $p(f_k = x_k | y = i)$by computing the attribute's mean and variance μ_{ik}, σ_{ik} from samples from that class C_i

Continuous Example

- Distinguish children from adults based on size
  - Classes $C = \{a, c\}$
  - Attributes: height[cm], weight[kg]
  - Training examples $\{h, w, y\}$, 4 adults, 12 children
- Class probabilities: $P(y = a)$?, $P(y = c)$?
- Model for adults:
  - Height ~Gaussian with
    - $\mu_{a,h} = \frac{1}{4}\sum_{i:y_i=a}(x_{i,h})$
    - $\sigma_{a,h}^2 = \frac{1}{4}\sum_{i:y_i=a}(x_{i,h} - \mu_{a,h})^2$
  - Weight ~ Gaussian $\mathcal{N}(\mu_{a,w}, \sigma_{a,w})$
- Model for children…
  - Height ~ $\mathcal{N}(\mu_{c,h}, \sigma_{c,h})$
  - Weight ~ $\mathcal{N}(\mu_{c,w}, \sigma_{c,w})$

Continuous Example

- Now given a test sample $x = (w, h)$ we want to compute $P(y = a|f = x)$ and $P(y = c|f = x)$
- To get our posteriors we want:
  - $P(y = a|f = x) = \frac{P(y=a)P(f = x|y = a)}{P(f=x)}$
  - $P(y = c|f = x) = \frac{P(y=c)P(f = x|y = c)}{P(f=x)}$
- Then if we make a naïve independence assumption we arrive at
  - $P(y = a|f = x) \propto P(y = a)P(f_1 = w|y = a)P(f_2 = h|y = a)$
  - $P(y = c|f = x) \propto P(y = c)P(f_1 = w|y = c)P(f_2 = h|y = c)$