Naïve Bayes

Prostate Cancer dataset – One field/class

It's useful to know:
$P(\text{cancer} = Y)$

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

Prostate Cancer dataset - One field/class

$P(C = Y)$ is $5/10 = 0.5$

P34 level	Prostate cancer
High	**Y**
Medium	**Y**
Low	**Y**
Low	N
Low	N
Medium	N
High	**Y**
High	N
Low	N
Medium	**Y**

So, with **no other info** you'd expect P(cancer=Y) to be 0.5

Prostate Cancer dataset - One field/class

But we know that P34 =H,
so actually we want:

***P*(*cancer=Y* | P34 = H)**

- the probability that cancer is Y, ***given that*** P34 is high

2/3 = 0.67

P34 level	**Prostate cancer**
High	**Y**
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	**Y**
High	N
Low	N
Medium	Y

Prostate Cancer dataset

Suppose again we know that P34 is High;

P (c=Y | **P34 = H**) = 0.5

P (c=N | **P34 = H**) = 0.25

P(c = Maybe | **H**) = 0.25

P34 level	**Prostate cancer**
High	**Y**
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	**Y**
High	N
High	Maybe
Medium	Y

Naive Bayes

$$P(\text{cancer} = \text{Y} \mid \mathbf{P34 = H})$$

P34 level	**Prostate cancer**
High	**Y**
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	**Y**
High	N
Low	N
Medium	Y

Naive Bayes

And now we are illustrating

$P(\text{P34} = \text{H} \mid \textbf{cancer = Y})$

2/5 = 0.4

P34 level	**Prostate cancer**
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

Naive Bayes

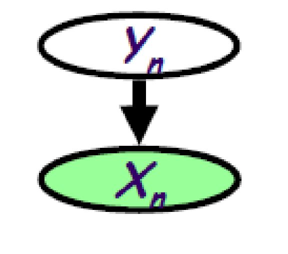
That is the essence of Naive Bayes,

but:

the probability calculations are much trickier when there are >1 fields

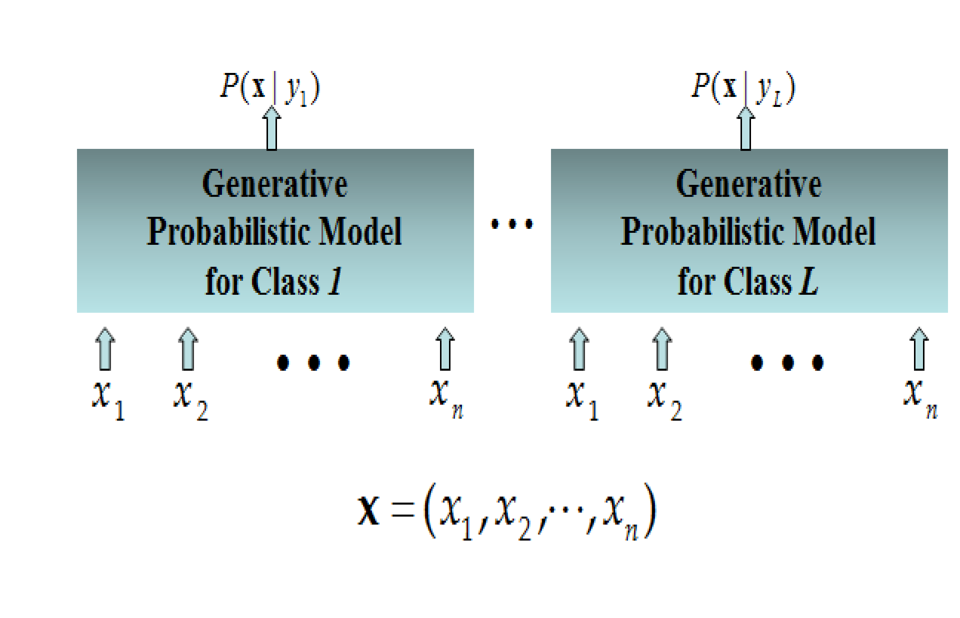
so we make a **'Naive'** assumption that makes it simpler

Probabilistic Classification Principle



- Establishing a probabilistic model for classification
  - **Generative model (must be probabilistic)**

$$P(\mathbf{x} \mid y) \quad y = y_1, \cdots, y_L, \ \mathbf{x} = (x_1, \cdots, x_n)$$



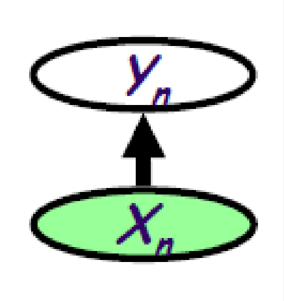
- *L* probabilistic models have to be trained independently
- Output *L* probabilities for a given input with *L* models
- Based on joint probability distribution

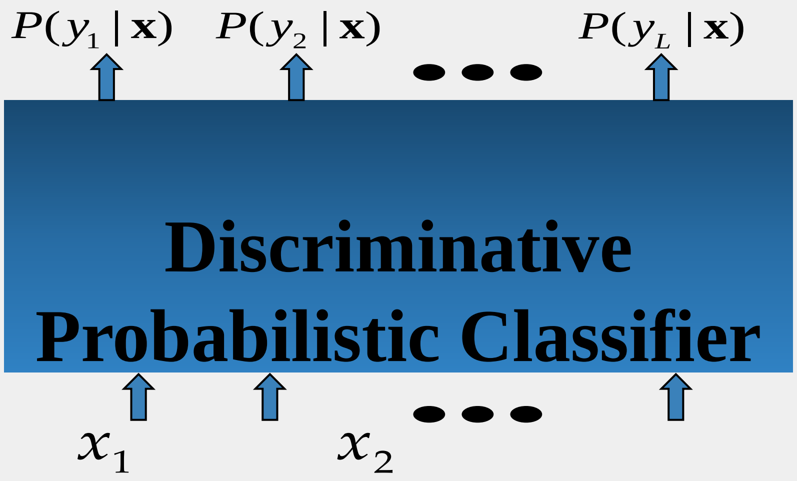
- Assume some functional form for P(X|Y), P(Y)
- Estimate parameters of P(X|Y), P(Y) directly from training data
- Use Bayes rule to calculate P(Y|X= x)

Probabilistic Classification Principle

- Establishing a probabilistic model for classification
  - **Discriminative (informative) model**

$$P(y \mid \mathbf{x}) \quad y = y_1, \cdots, y_L, \ \mathbf{x} = (x_1, \cdots, x_n)$$





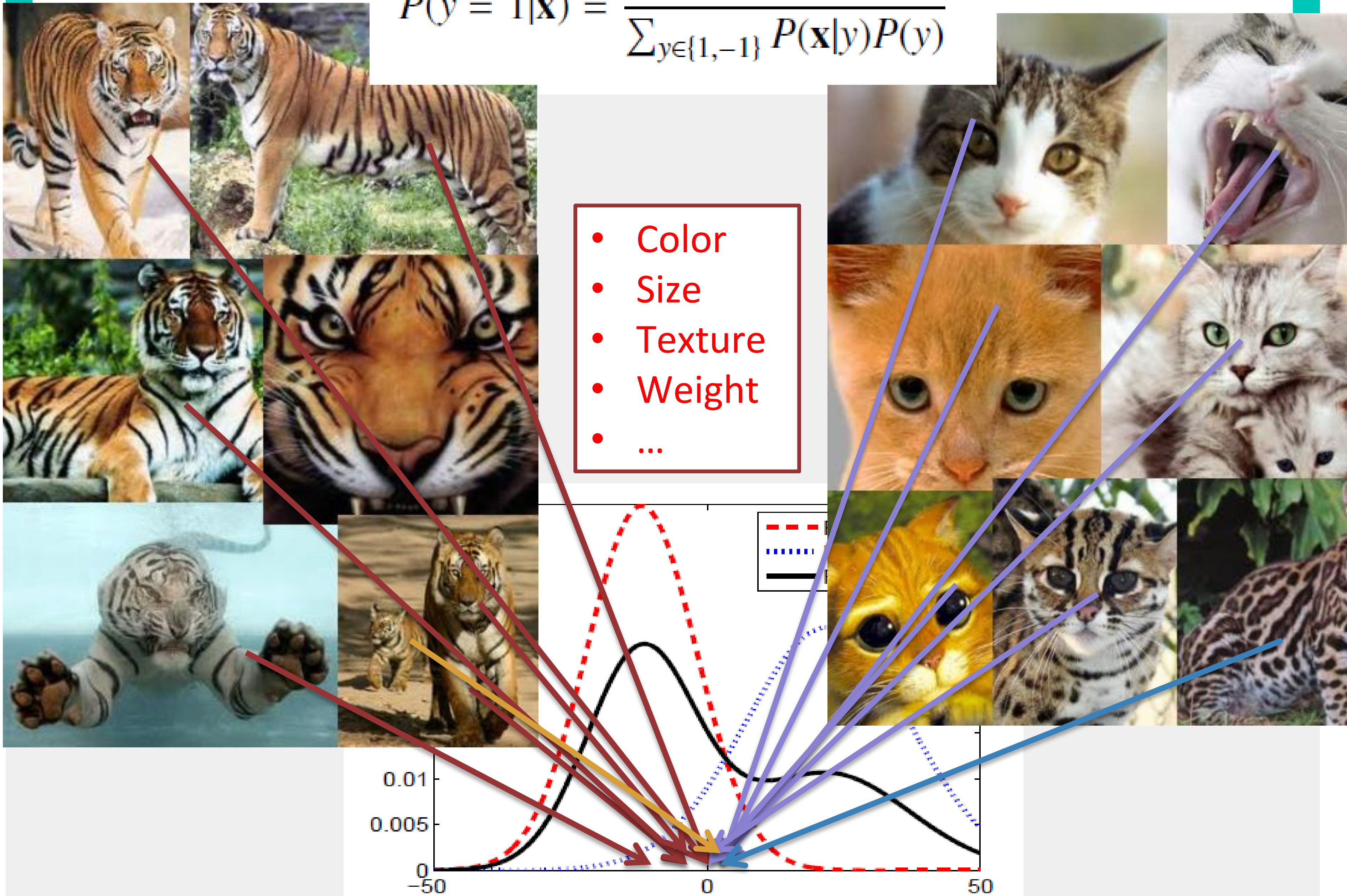
$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

- To train a discriminative classifier regardless its probabilistic or non-probabilistic nature, all training examples of different classes must be jointly used to build up a single discriminative classifier.
- Directly assume some functional form for P(Y|X)
- Estimate parameters of P(Y|X) directly from training data

Generative Mod

$$P(y=1|\mathbf{x}) = \frac{P(\mathbf{x}|y=1)P(y=1)}{\sum_{y\in\{1,-1\}} P(\mathbf{x}|y)P(y)}$$

- Color
- Size
- Texture
- Weight
- …



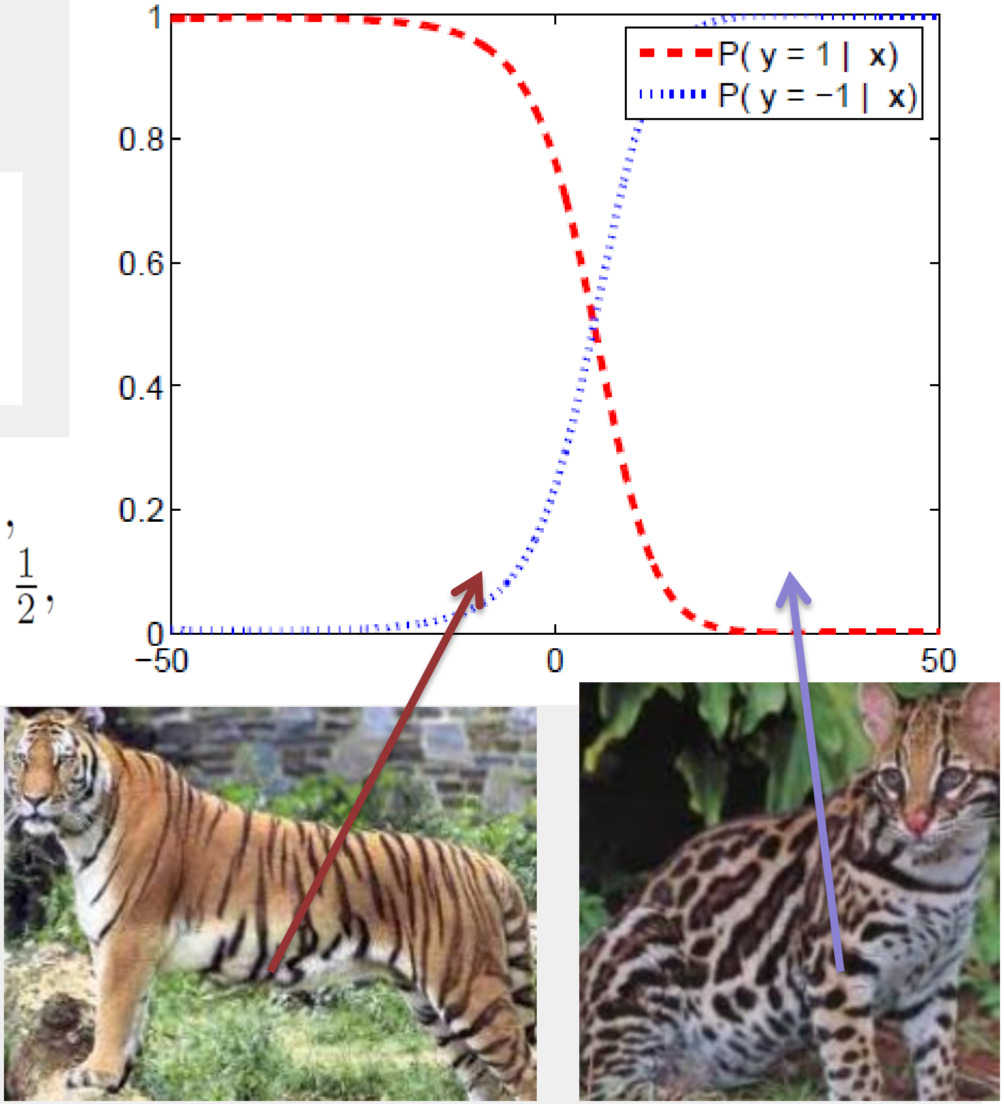
Discriminative Model

- Logistic Regression

$$P(y=1|\mathbf{x}) = \frac{1}{1+\exp(yf(\mathbf{x}))}$$

$$f^*(\mathbf{x}) = \begin{cases} +\infty & \Pr(y=1|\mathbf{x}) > \frac{1}{2}, \\ -\infty & \Pr(y=-1|\mathbf{x}) < \frac{1}{2}, \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

- Color
- Size
- Texture
- Weight
- …



Bayes Formula

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Likelihood of seeing the evidence if the hypothesis is correct

Posterior probability

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

Prior probability

Normalizing constant – the likelihood of the evidence under any circumstances

Bayes Theorem – Additional Info

- Prior, conditional and joint probability for random variables
  - Prior probability: $P(X)$
  - Conditional probability: $P(X_1 \mid X_2), P(X_2 \mid X_1)$
  - Joint probability: $\mathbf{X} = (X_1, X_2), P(\mathbf{X}) = P(X_1, X_2)$
  - Relationship: $P(X_1, X_2) = P(X_2 \mid X_1)P(X_1) = P(X_1 \mid X_2)P(X_2)$
  - Independence: $P(X_2 \mid X_1) = P(X_2), P(X_1 \mid X_2) = P(X_1), P(X_1, X_2) = P(X_1)$

Bayes Theorem

Bayes theorem deals with ***sequential events,*** *whereby* ***new*** additional information is obtained for a subsequent event, and that new information is used to revise the probability of the initial event.

Prior probability - is an initial probability value originally obtained before any additional information is obtained.

Posterior probability - is a probability value that has been revised by using additional information that is later obtained.

Bayes Theorem – Example #1

An organization randomly selects an adult for a survey about credit card usage. Use subjective probabilities to estimate the following.

a. What is the probability that the selected subject is a male?

b. After selecting a subject, it is later learned that this person was smoking a cigar during the interview. What is the probability that the selected subject is a male?

c. Which of the preceding two results is a prior probability? Which is a posterior probability?

Bayes Theorem - Example

Roughly half of all population are males, so we estimate the probability of selecting a male subject to be 0.5. Denoting a male by M, we can express this probability as follows: ***P(M) = 0.5.***

b. Although some women smoke cigars, the vast majority of cigar smokers are males. A reasonable guess is that 85% of cigar smokers are males. Based on this additional subsequent information that the survey respondent was smoking a cigar, we estimate the probability of this person being a male as 0.85. Denoting a male by M and denoting a cigar smoker by C, we can express this result as follows: ***P(M | C) = 0.85***.

c. In part (a), the value of 0.5 is the initial probability, so we refer to it as the prior probability. Because the probability of 0.85 in part (b) is a revised probability based on the additional information that the survey subject was smoking a cigar, this value of 0.85 is referred to a posterior probability.

Bayes Theorem – Example #2

Now assume that in District A, 51% of the adults are males. One adult is randomly selected for a survey involving credit card usage.

a. Find the prior probability that the selected person is a male.

b. It is later learned that the selected survey subject was smoking a cigar. Also, 9.5% of males smoke cigars, whereas 1.7% of females smoke cigars.

Use this additional information to find the **probability that the selected subject is a male.**

Bayes Theorem - Example

Let's use the following notation:

M = male M' = female (or not male)

C = cigar smoker C'= not a cigar smoker.

a. The probability of randomly selecting an adult and getting a male is given by ***P(M) = 0.51.***

b. Based on the additional given information, we have the following:

P(M) = 0.51 *because 51% of the adults are males*

P(M') = 0.49 *because 49% of the adults are females (not males)*

P(C|M) = 0.095 *because 9.5% of the males smoke cigars*

P(C|M') = 0.017. *because 1.7% of the females smoke cigars*

Bayes Theorem - Example

$$P(M \mid C) = \frac{P(M) \cdot P(C|M)}{[P(M) \cdot P(C|M)] + [P(\overline{M}) \cdot P(C|\overline{M})]}$$

$$= \frac{0.51 \cdot 0.095}{[0.51 \cdot 0.095] + [0.49 \cdot 0.017]}$$

$$= 0.85329341$$

$$= 0.853 \text{ (rounded)}$$

Initially we knew that the survey subject smoked a cigar, there is a 0.51 probability that the survey subject is male, however, after learning that the subject smoked a cigar, we revised the probability to 0.853.

#Checkpoint

- Given:
  - A doctor knows that meningitis (M) causes stiff neck (S) 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis?

#Checkpoint

There is a school with 60% boys and 40% girls. The girls wear trousers or skirts in equal numbers; all the boys wear trousers. An observer sees a student wearing trousers. What is the probability this student is a girl?

#Checkpoint

While watching a game of football in a cafe, you observe someone who is clearly supporting Manchester United in the game. **What is the probability that they were actually born within 25 miles of Manchester?** Assume that:

- the probability that a randomly selected person is born within 25 miles of Manchester is 1/20;

- the chance that a person born within 25 miles of Manchester actually supports United is 7/10;

- the probability that a person not born within 25 miles of Manchester supports United with probability 1/10

Naïve Bayes with Many Fields

New patient:
P34=M, P61=M, BMI = H

Best guess at cancer field ?

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

Naïve Bayes with Many Fields

$$P(\text{p34=M} \mid \text{Y}) \times P(\text{p61=M} \mid \text{Y}) \times P(\text{BMI=H} \mid \text{Y}) \times P(\text{cancer} = \text{Y})$$
$$P(\text{p34=M} \mid \text{N}) \times P(\text{p61=M} \mid \text{N}) \times P(\text{BMI=H} \mid \text{N}) \times P(\text{cancer} = \text{N})$$

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

Naïve Bayes with Many Fields

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	**Y**
Medium	Low	Medium	**Y**
Low	Low	High	**Y**
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	**Y**
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	**Y**

$$0.4 \times 0 \times 0.4 \times 0.5 = 0$$

$$0.2 \times 0.4 \times 0.2 \times 0.5 = 0.008$$

Naïve Bayes with Many Fields

In practice, we finesse the zeroes and use logs:
(note: $\log(A \times B \times C \times D \times \ldots) = \log(A) + \log(B) + \ldots$)

$\log(0.4) + \log(0.001) + \log(0.4) + \log(0.5) = -4.09$
$\log(0.2) + \log(0.4) + \log(0.2) + \log(0.5) = -2.09$

Naïve Bayes in General

Essence of Naive Bayes,   with 1 non-class field, is to calculate this for each class value, given some new instance with field = F:

P(class = C | Field = F)

For many fields, our new instance is (e.g.) (F1, F2, ...Fn), and the 'essence of Naive Bayes' is to calculate *this* for each class:

P(class = C | F1,F2,F3,...,Fn)

i.e. What is probability of class C, given all these field values together?

Naïve Bayes in General

- Naïve Bayes Algorithm (for discrete input attributes)
  - Learning Phase: Given a training set **S**,

$$
\begin{aligned}
&\text{For each target value of } c_i \ (c_i = c_1, \cdots, c_L) \\
&\quad \hat{P}(C = c_i) \leftarrow \text{estimate } P(C = c_i) \text{ with examples in } \mathbf{S}; \\
&\quad \text{For every attribute value } x_{jk} \text{ of each attribute } X_j \ (j = 1, \cdots, n;\ k = 1, \cdots, N_j) \\
&\qquad \hat{P}(X_j = x_{jk} \mid C = c_i) \leftarrow \text{estimate } P(X_j = x_{jk} \mid C = c_i) \text{ with examples in } \mathbf{S};
\end{aligned}
$$

    Output: conditional probability tables; for X_j, $N_j \times L$ elements

  - Test Phase: Given an unknown instance $\mathbf{X}' = (a'_1, \cdots, a'_n)$

    Look up tables to assign the label c^* to $\mathbf{X}'$ if

$$
[\hat{P}(a'_1 \mid c^*) \cdots \hat{P}(a'_n \mid c^*)]\hat{P}(c^*) > [\hat{P}(a'_1 \mid c) \cdots \hat{P}(a'_n \mid c)]\hat{P}(c), \quad c \neq c^*, c = c_1, \cdots, c_L
$$

Naïve Bayes - Example

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Naïve Bayes - Example

- Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=Yes) = 9/14 \quad P(\text{Play}=No) = 5/14$

Naïve Bayes – Continuous Features

- Algorithm: Continuous-valued Features
  - Numberless values taken by a continuous-valued feature
  - Conditional probability often modeled with the normal distribution

$$\hat{P}(x_j \mid y_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(x_j - \mu_{ji})^2}{2\sigma_{ji}^2} \right)$$

μ_{ji} : mean (average) of feature values x_j of examples for which $y = y_i$

σ_{ji} : standard deviation of feature values x_j of examples for which $y = y_i$

  - Learning Phase: for $\mathbf{X} = (X_1, \cdots, X_F)$, $\mathbf{Y} = y_1, \cdots, y_L$

    Output: $F \times L$ normal distributions and $P(Y = y_i)$ $i = 1, \cdots, L$
  - Test Phase: Given an unknown instance $\mathbf{X}' = (a'_1, \cdots, a'_n)$
    - Instead of looking-up tables, calculate conditional probabilities with all the normal distributions achieved in the learning phrase
    - Apply the Maximum A Posteriori rule to assign a label (the same as done for the discrete case)

Naïve Bayes – Continuous Features

- Example: Continuous-valued Features
  - Temperature is naturally of continuous value.

    Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

    No: 27.3, 30.1, 17.4, 29.5, 15.1
  - Estimate mean and variance for each class

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \ \sigma_{Yes} = 2.35$$
$$\mu_{No} = 23.88, \ \sigma_{No} = 7.09$$

  - **Learning Phase**: output two Gaussian models for P(temp|Y)

$$\hat{P}(x \mid Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{2 \times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$\hat{P}(x \mid No) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{2 \times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$

Conclusion

- Probabilistic Classification Principle
  Discriminative vs. Generative models: learning P(y|x) vs. P(x|y)

- Naïve Bayes: the conditional independence assumption
  Working well sometimes for data violating the assumption!

- A popular generative model
  Performance competitive to most of state-of-the-art classifiers even in presence of violating independence assumption
  Many successful applications, e.g., spam mail filtering
  A good candidate of a **base learner** in ensemble learning