Attendance Code:
xxxx

Artificial Intelligence and Machine Learning (AIML)

2023–24

#Code

- **Last lecture**: probability, probabilistic graphical models

Given two (or more) random variables, X and Y, with their corresponding PMFs/PDFs,

Joint Probability:

$$P(X,Y) = P(X=x, Y=y)$$

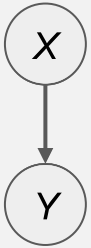
Marginal Probability:

$$P(X=x) = \sum_{y \in \Omega_Y} P(X=x, Y=y)$$

Conditional Probability:

$$P(X,Y) = P(Y|X)P(X)$$

Probabilistic Graphical Model (PGM)



- **Last lecture**: probability, probabilistic graphical models

Given two (or more) random variables, X and Y, with their corresponding PMFs/PDFs,

Joint Probability:

$$P(X,Y) = P(X=x, Y=y) \underset{\text{independent}}{=} P(X=x)P(Y=y)$$

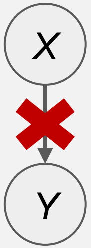
Marginal Probability:

$$P(X=x) = \sum_{y \in \Omega_Y} P(X=x, Y=y)$$

Conditional Probability:

$$P(X,Y) = P(Y|X)P(X) \xrightarrow{\text{independent}} P(Y|X) = P(Y)$$

Probabilistic Graphical Model (PGM)



#Code

- **Last lecture**: probability, probabilistic graphical models

Given two (or more) random variables, X and Y, with their corresponding PMFs/PDFs,

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X,U,V,Z)$$

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(U,V,Z)$$

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(Z|U,V)P(U,V)$$

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(Z|U,V)P(U|V)P(V)$$

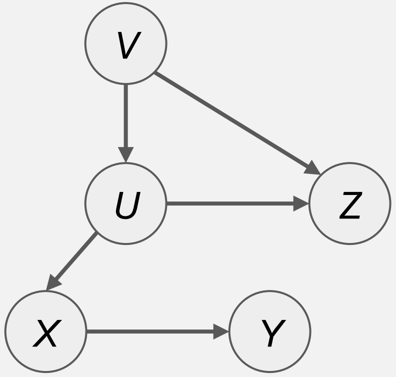
Simplify using conditional independence:

$$P(U,V,X,Y,Z) = P(Y|X,U,V,Z)P(X|U,V,Z)P(Z|U,V)P(U|V)P(V)$$

$$P(U,V,X,Y,Z) = P(V)P(U|V)P(X|U)P(Y|X)P(Z|U,V)$$

Markov factorization

Probabilistic Graphical Model (PGM)



- **Last lecture**: probability, probabilistic graphical models

- **This lecture**: How to use probability in classification

(Contra)Intuitive Example

Hypothetical Situation:

You wake up and feel sick. You go to the doctor and have a test taken. After a week goes by, the results come back, and it turns out you tested positive for a rare disease that only affects 0.1% of the population.

Your doctor says that the test correctly identifies 99% of people who have the disease and only incorrectly identifies 1% of people who don’t have the disease.

How concerned would you be? What are the chances that you do have this disease?

Bayes' theorem

- We learned that $P(X,Y) = P(Y|X)P(X)$.
- However, $P(Y,X) = P(X|Y)P(Y)$ should be equivalent to $P(X,Y)$

- The relations above allow us to swap conditionals:

likelihood prior

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

(Bayes' Theorem)

posterior evidence

- If evidence distribution $P(Y)$ is unknown, can use instead:

$$P(Y) = \sum_{x \in \Omega_X} P(Y|X = x)P(X = x)$$

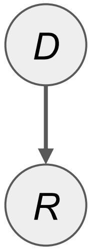
#Code

Bayes' theorem provides a rational synthesis of uncertainty

- **Problem**: determining disease status given a test result
- **Distributions**:
  - R - result (Bernoulli, $\Omega_R = \{0,1\}$)
  - D - health status (Bernoulli, $\Omega_D = \{h,d\}$ for 'healthy' and 'disease', respectively)
- **Likelihood data**: from observations, $P(R = 1|D = d) = 0.99$, $P(R = 1|D = h) = 0.01$
- **Prior data**: (often ignored in "standard" reasoning, but can be considered from medical literature), $P(D = d) = 0.001$
- **Graphical model**: result depends on health state,

$$P(R, D) = P(R|D)P(D)$$

Read lecture notes for a problem using a categorical distribution (N = 3)



#Code

Bayes' theorem provides a rational synthesis of uncertainty

- **Bayes' theorem**: posterior probability of each health state, given the test result,

$$P(D|R=1)=\frac{P(R=1|D)P(D)}{P(R=1)}$$

- **Evidence unknown**, so must **marginalize**:

$$P(R=1)=\sum_{D\in\Omega_D}P(R|D)P(D)=P(R=1|D=d)P(D=d)+P(R=1|D=h)P(D=h)$$

- **Calculations** using data are:

$$P(D|R=1)=\frac{0.99\times 0.001}{0.99\times 0.001+0.01\times 0.999}=0.09$$

For 0.1% people have the disease, 99% have been correctly identified with the test result.
For 99.9% people don't have the disease, but only 1% has been incorrectly identified.

AIML

#Code

Bayes' theorem provides a rational synthesis of uncertainty

For 1000 people, 990 identified correctly.
For 1000, 10 incorrectly identified with positively.

So the total number of R=1 is 11: 1 + 10 (actual have disease + false positive)

For example, we have 1000 people

Health status	***Prior*** $P(D)$	***Likelihood*** $P(R{=}1 \mid D)$	***Posterior*** $P(D \mid R{=}1)$
$D{=}h$	999 0.999	0.99	0.91
$D{=}d$	1 0.001	0.01	0.09

10/11, among the positive test results (R=1), the vast majority (91%) were false positives

1/11 (approximately 9%): This represents the positive predictive value (PPV)

- **Conclusion**: after having the test result, being healthy is still the most probable status, but having the rare disease has gone from 0.1% probability to 9%, should not be ignored in this situation
- **Bayes'** is **precise synthesis** of disparate sources of **uncertain information**

So for a person who has the first test R=1, he/she might be lucky that he/she is in one of those 10 people who are falsely identified with positive R=1.

AIML

#Code

Bayes' theorem provides a rational synthesis of uncertainty

Health status	***Prior*** $P(D)$	***Likelihood*** $P(R=1\|D)$	***Posterior*** $P(D\|R=1)$
$D=h$	0.999	0.99	0.91
$D=d$	0.001	0.01	0.09

- If we run a second, independent test, after having tested positive for the first result

$$P(D|R=1) = \frac{P(R=1|D)P(D)}{P(R=1)} = \frac{0.99 \times 0.09}{0.99 \times 0.09 + 0.01 \times 0.91} = 0.91$$

#Code

Probabilistic classification using Bayes' theorem

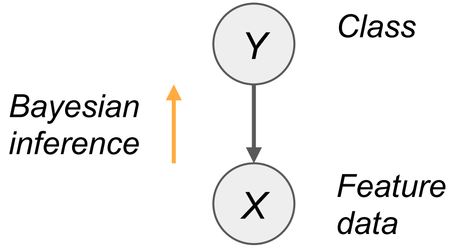
- **Probabilistic classification** can be expressed as an application of Bayes' rule:
  - given some **input** (feature) data X, determine the **probability** $P(Y|X)$ of the **class Y to which X belongs** (posterior), taking into account $P(Y)$ (prior) and how probable that class is before having seen the data, $P(X|Y)$

- A good decision is to select the value of Y which maximizes $P(Y|X)$, called the **maximum a-posteriori** (MAP) decision:

$$y^\star = \arg\max_{y \in \Omega_Y} P(Y = y | X = x)$$

and can avoid the need to have the evidence $P(X = x)$, since it does not depend upon Y:

$$y^\star = \arg\max_{y \in \Omega_Y} P(X = x | Y = y) P(Y = y)$$

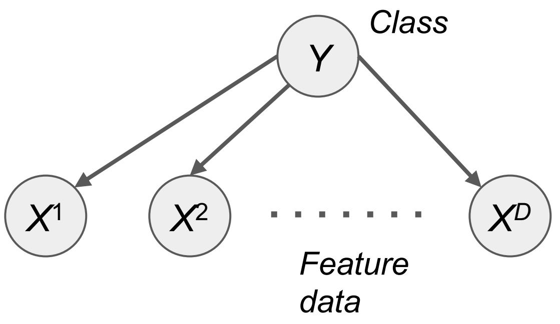


AIML

Naive Bayes classifier: MAP solution

#Code

- In general, the input features X will be **multidimensional** (a vector of values) and will not be independent of each other making it difficult to estimate the likelihood $P(X|Y)$ from the data
- The so-called **naive Bayes' classifier** simplifies the classification model by assuming that each feature is conditionally independent of the others, given the class.



Markov Factorization:

$$P(X|Y) = P(X^1|Y)P(X^2|Y)\cdots P(X^D|Y)$$

Using Bayes' theorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X^1|Y)P(X^2|Y)\cdots P(X^D|Y)P(Y)}{P(X)}$$

Since $P(X)$ is independent of Y,

$$y^\star = \arg\max_{y\in\Omega_Y} P(X^1|Y)P(X^2|Y)\cdots P(X^D|Y)P(Y=y)$$

#Code

Naive Bayes classifier: example

- **Problem**: spam detection
- **Likelihood feature distributions**:
  - $D = 4$ features (X^1, X^2, X^3, X^4)
    - words in the email (dear, friend, thank, buy)
  - Class labels: Ω_Y={'S', 'R'} (spam, not spam or regular)
- **Class priors**:
  - from training data: 15 emails (10 regular, 5 spam): $P(R) = 2/3$, $P(S) = 1/3$

Regular emails:

- dear: 8 out of 17 words - $P(X^1|Y = R) = \frac{8}{17} = 0.47$
- friend: 5 out of 17 words - $P(X^2|Y = R) = \frac{5}{17} = 0.29$
- thank: 3 out of 17 words - $P(X^3|Y = R) = \frac{3}{17} = 0.18$
- buy: 1 out of 17 words - $P(X^4|Y = R) = \frac{1}{17} = 0.06$

Spam emails:

- dear: 4 out of 17 words - $P(X^1|Y = S) = \frac{\overset{4}{\cancel{5}}}{17} = 0.24$
- friend: 2 out of 17 words - $P(X^2|Y = S) = \frac{2}{17} = 0.12$
- thank: 1 out of 17 words - $P(X^3|Y = R) = \frac{1}{17} = 0.06$
- buy: 10 out of 17 words - $P(X^4|Y = R) = \frac{10}{17} = 0.59$

AIML

#Code

Naive Bayes classifier: example

- **New email** containing words “friend” and “thank”
- Is it likely to be a regular email or spam? $y^\star = \underset{y \in \Omega_Y}{\arg\max}\, P(X^2|Y)P(X^3|Y)P(Y=y)$
  - $Y = R$: $\mathrm{p}(\mathrm{Y} = R|X) = 0.29 \times 0.18 \times 0.67 = 0.035$
  - $Y = S$: $\mathrm{p}(\mathrm{Y} = S|X) = 0.12 \times 0.06 \times 0.33 = 0.002$
- $y^\star = R$: **new email is likely NOT to be a spam**

Note that the result is the same regardless of the order of “dear” and “friend” in the email

  - from training data: 15 emails (10 regular, 5 spam): $P(R) = 2/3$, $P(S) = 1/3$

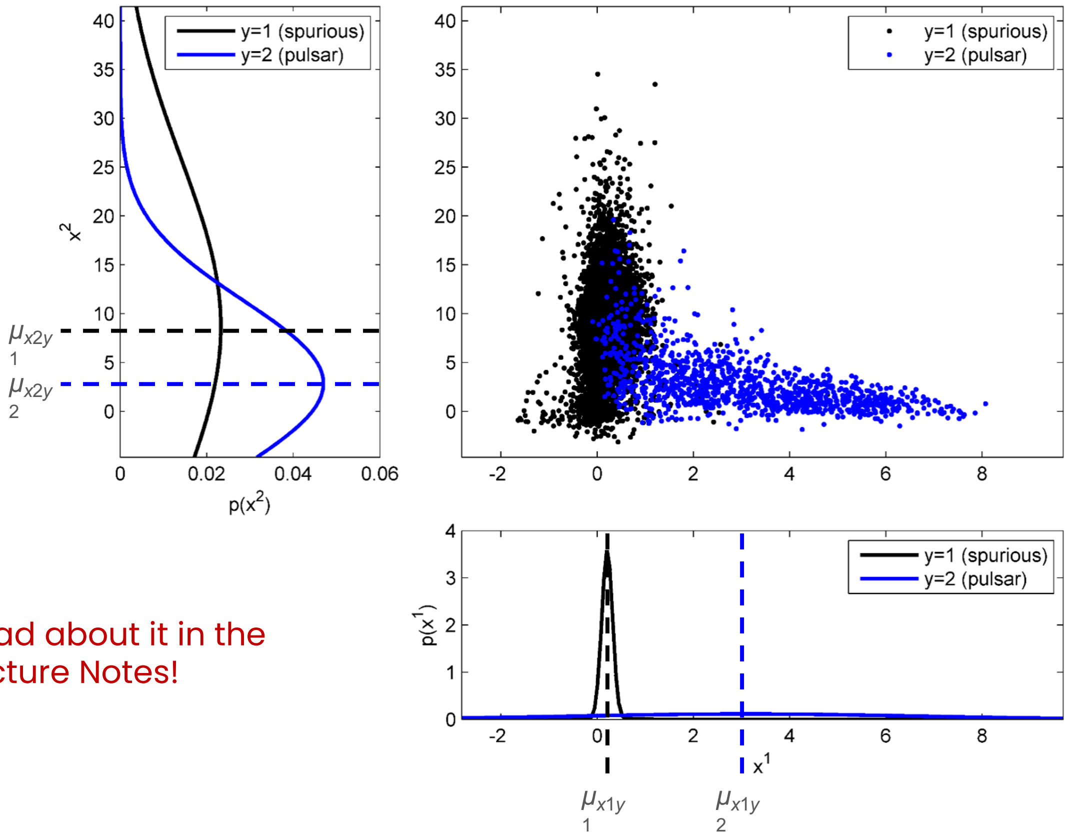
Regular emails:
- dear: 8 out of 17 words - $P(X^1|Y = R) = \frac{8}{17} = 0.47$
- friend: 5 out of 17 words - $P(X^2|Y = R) = \frac{5}{17} = 0.29$
- thank: 3 out of 17 words - $P(X^3|Y = R) = \frac{3}{17} = 0.18$
- buy: 1 out of 17 words - $P(X^4|Y = R) = \frac{1}{17} = 0.06$

Spam emails:
- dear: 4 out of 17 words - $P(X^1|Y = S) = \frac{5}{17} = 0.24$
- friend: 2 out of 17 words - $P(X^2|Y = S) = \frac{2}{17} = 0.12$
- thank: 1 out of 17 words - $P(X^3|Y = R) = \frac{1}{17} = 0.06$
- buy: 10 out of 17 words - $P(X^4|Y = R) = \frac{10}{17} = 0.59$

AIML

#Code

Naive Bayes classifier in Astrophysics



Read about it in the Lecture Notes!

- On test data, compute most probable class for each case: y^*_i for $i = 1,2,\ldots, N_{\text{test}}$
- Compute **0-1 error function** using known test labels y_i
- Test error: **~80% correctly identified**
- Use posterior probabilities to check only **uncertain** decisions (<10% of total)

Naive Bayes classifier: analysis

- Naive Bayes surprisingly good for high-dimensional problems (D large), since **does not require a large amount of training data**
- **Estimating feature distribution** parameters is **very quick**: linear in D, the number of features
- **Making a prediction** requires **evaluating D times** $|\Omega_Y|$ (the number of classes), which is usually easy to carry out in practice
- Nonetheless, assumption of **feature independence is unrealistic** for many practical ML problems

AIML

#Code

To recap

- We discussed how we can use Bayes' theorem for classification problems
  - **Naïve Bayes' classifier**: quite efficient; assumes features are independent
- We learned how to make predictions using the naïve Bayes' classifier
- **Next**: Sequence modelling and hidden Markov models

Further Reading

- **PRML**, Section 1.2
- **R&N**, Sections 21.1 and 21.2
- **MLSP**, Section 1.4