INF367 25H: Selected Topics in Artificial Intelligence

Diamonds and Rust in the AI Treasure Chest

Dario Garigliotti - UiB - 19.09.2025

Plan for today

- Recap from last lecture
- Primer on probability
    - Basics, including Bayes rule
    - Example: Monty Hall
    - Activity 1
    - MLE, MAP
    - Bayesian learning
    - Activity 2

Probability

Probability

- Basic notions
  - Definition
  - Experiment, event
  - Random variable
  - Distribution, parameters
  - Conditional, marginal
  - Bayes

Probability

Sets

Definition 1.2 (Set notation). An alternative notation in terms of set theory is to write

$$p(x \text{ or } y) \equiv p(x \cup y), \qquad p(x, y) \equiv p(x \cap y)$$

Probability

Marginal

Definition 1.3 (Marginals)**.** Given a *joint distribution* $p(x, y)$ the distribution of a single variable is given

$$p(x) = \sum_{y} p(x, y)$$

Probability

Conditional, Bayes

Definition 1.4 (Conditional Probability / Bayes' Rule)**.** The probability of event x conditioned on knowing event y (or more shortly, the probability of x given y) is defined as

$$p(x|y) \equiv \frac{p(x, y)}{p(y)} \tag{1.1.7}$$

If $p(y) = 0$ then $p(x|y)$ is not defined. From this definition and $p(x, y) = p(y, x)$ we immediately arrive at Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{1.1.8}$$

Activity 1

Activity 1

Example 1.2 (Hamburgers). Consider the following fictitious scientific information: Doctors find that people with Kreuzfeld-Jacob disease (KJ) almost invariably ate hamburgers, thus $p(\mathit{Hamburger\ Eater}|KJ) = 0.9$. The probability of an individual having KJ is currently rather low, about one in 100,000.

1. Assuming eating lots of hamburgers is rather widespread, say $p(\mathit{Hamburger\ Eater}) = 0.5$, what is the probability that a hamburger eater will have Kreuzfeld-Jacob disease?

2. If the fraction of people eating hamburgers was rather small, $p(\mathit{Hamburger\ Eater}) = 0.001$, what is the probability that a regular hamburger eater will have Kreuzfeld-Jacob disease?

Activity 1

Example 1.2 (Hamburgers)**.** Consider the following fictitious scientific information: Doctors find that people with Kreuzfeld-Jacob disease (KJ) almost invariably ate hamburgers, thus $p(Hamburger\ Eater|KJ\ ) = 0.9$. The probability of an individual having KJ is currently rather low, about one in 100,000.

1. Assuming eating lots of hamburgers is rather widespread, say $p(Hamburger\ Eater) = 0.5$, what is the probability that a hamburger eater will have Kreuzfeld-Jacob disease?

This may be computed as

$$p(KJ\ |Hamburger\ Eater) = \frac{p(Hamburger\ Eater, KJ\ )}{p(Hamburger\ Eater)} = \frac{p(Hamburger\ Eater|KJ\ )p(KJ\ )}{p(Hamburger\ Eater)} \tag{1.2.1}$$

$$p(x|y) \equiv \frac{p(x,y)}{p(y)}$$

$$= \frac{\frac{9}{10} \times \frac{1}{100000}}{\frac{1}{2}} = 1.8 \times 10^{-5} \tag{1.2.2}$$

Activity 1

Example 1.2 (Hamburgers)**.** Consider the following fictitious scientific information: Doctors find that people with Kreuzfeld-Jacob disease (KJ) almost invariably ate hamburgers, thus $p(Hamburger\ Eater|KJ\ ) = 0.9$. The probability of an individual having KJ is currently rather low, about one in 100,000.

2. If the fraction of people eating hamburgers was rather small, $p(Hamburger\ Eater) = 0.001$, what is the probability that a regular hamburger eater will have Kreuzfeld-Jacob disease? Repeating the above calculation, this is given by

$$\frac{\frac{9}{10} \times \frac{1}{100000}}{\frac{1}{1000}} \approx 1/100 \tag{1.2.3}$$

   This is much higher than in scenario (1) since here we can be more sure that eating hamburgers is related to the illness.

Bayesian Learning

Probability

Marginal

Definition 1.3 (Marginals)**.** Given a *joint distribution* $p(x, y)$ the distribution of a single variable is given

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

Probability

Conditional, Bayes

Definition 1.4 (Conditional Probability / Bayes' Rule)**.** The probability of event x conditioned on knowing event y (or more shortly, the probability of x given y) is defined as

$$p(x|y) \equiv \frac{p(x, y)}{p(y)} \tag{1.1.7}$$

If $p(y) = 0$ then $p(x|y)$ is not defined. From this definition and $p(x, y) = p(y, x)$ we immediately arrive at Bayes' rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{1.1.8}$$

Naive Bayes

$$p(\mathbf{x}, c) = p(c) \prod_{i=1}^{D} p(x_i|c)$$

$$p(c|\mathbf{x}^*) = \frac{p(\mathbf{x}^*|c)p(c)}{p(\mathbf{x}^*)} = \frac{p(\mathbf{x}^*|c)p(c)}{\sum_c p(\mathbf{x}^*|c)p(c)}$$

Naive Bayes

$$p(\mathbf{x}, c) = p(c) \prod_{i=1}^{D} p(x_i|c)$$

$$p(c|\mathbf{x}^*) = \frac{p(\mathbf{x}^*|c)p(c)}{p(\mathbf{x}^*)} = \frac{p(\mathbf{x}^*|c)p(c)}{\sum_c p(\mathbf{x}^*|c)p(c)}$$

$$p(x) = \sum_y p(x, y) \qquad p(x|y) \equiv \frac{p(x, y)}{p(y)}$$

Activity 2

Activity 2

Example 10.1. EZsurvey.org partitions radio station listeners into two groups – the 'young' and 'old'. They assume that, given the knowledge that a customer is either 'young' or 'old', this is sufficient to determine whether or not a customer will like a particular radio station, independent of their likes or dislikes for any other stations:

$$p(r_1, r_2, r_3, r_4|age) = p(r_1|age)p(r_2|age)p(r_3|age)p(r_4|age) \tag{10.1.3}$$

where each of the variables r_1, r_2, r_3, r_4 can take the states like or dislike, and the 'age' variable can take the value young or old. Thus the information about the age of the customer determines the individual radio station preferences without needing to know anything else. To complete the specification, given that a customer is young, she has a 95% chance to like Radio1, a 5% chance to like Radio2, a 2% chance to like Radio3 and a 20% chance to like Radio4. Similarly, an old listener has a 3% chance to like Radio1, an 82% chance to like Radio2, a 34% chance to like Radio3 and a 92% chance to like Radio4. They know that 90% of the listeners are old.

Given this model, and the fact that a new customer likes Radio1, and Radio3, but dislikes Radio2 and Radio4, what is the probability that the new customer is young?

Activity 2

Given this model, and the fact that a new customer likes Radio1, and Radio3, but dislikes Radio2 and Radio4, what is the probability that the new customer is young? This is given by

$$p(\mathsf{young}|r_1 = \mathsf{like}, r_2 = \mathsf{dislike}, r_3 = \mathsf{like}, r_4 = \mathsf{dislike})$$
$$= \frac{p(r_1 = \mathsf{like}, r_2 = \mathsf{dislike}, r_3 = \mathsf{like}, r_4 = \mathsf{dislike}|\mathsf{young})p(\mathsf{young})}{\sum_{age} p(r_1 = \mathsf{like}, r_2 = \mathsf{dislike}, r_3 = \mathsf{like}, r_4 = \mathsf{dislike}|age)p(age)} \tag{10.1.4}$$

radio station preferences without needing to know anything else. To complete the specification, given that a customer is young, she has a 95% chance to like Radio1, a 5% chance to like Radio2, a 2% chance to like Radio3 and a 20% chance to like Radio4. Similarly, an old listener has a 3% chance to like Radio1, an 82% chance to like Radio2, a 34% chance to like Radio3 and a 92% chance to like Radio4. They know that 90% of the listeners are old.

Activity 2

Given this model, and the fact that a new customer likes Radio1, and Radio3, but dislikes Radio2 and Radio4, what is the probability that the new customer is young? This is given by

$$p(\text{young}|r_1 = \text{like}, r_2 = \text{dislike}, r_3 = \text{like}, r_4 = \text{dislike}) \\ = \frac{p(r_1 = \text{like}, r_2 = \text{dislike}, r_3 = \text{like}, r_4 = \text{dislike}|\text{young})p(\text{young})}{\sum_{age} p(r_1 = \text{like}, r_2 = \text{dislike}, r_3 = \text{like}, r_4 = \text{dislike}|age)p(age)} \tag{10.1.4}$$

Using the naive Bayes structure, the numerator above is given by

$$p(r_1 = \text{like}|\text{young})p(r_2 = \text{dislike}|\text{young})p(r_3 = \text{like}|\text{young})p(r_4 = \text{dislike}|\text{young})p(\text{young}) \tag{10.1.5}$$

Plugging in the values we obtain

Activity 2

Given this model, and the fact that a new customer likes Radio1, and Radio3, but dislikes Radio2 and Radio4, what is the probability that the new customer is young? This is given by

$$p(\mathsf{young}|r_1 = \mathsf{like}, r_2 = \mathsf{dislike}, r_3 = \mathsf{like}, r_4 = \mathsf{dislike})$$
$$= \frac{p(r_1 = \mathsf{like}, r_2 = \mathsf{dislike}, r_3 = \mathsf{like}, r_4 = \mathsf{dislike}|\mathsf{young})p(\mathsf{young})}{\sum_{age} p(r_1 = \mathsf{like}, r_2 = \mathsf{dislike}, r_3 = \mathsf{like}, r_4 = \mathsf{dislike}|age)p(age)} \tag{10.1.4}$$

The denominator is given by this value plus the corresponding term evaluated assuming the customer is old,

$$0.03 \times 0.18 \times 0.34 \times 0.08 \times 0.9 = 1.3219 \times 10^{-4}$$

radio station preferences without needing to know anything else. To complete the specification, given that a customer is young, she has a 95% chance to like Radio1, a 5% chance to like Radio2, a 2% chance to like Radio3 and a 20% chance to like Radio4. Similarly, an old listener has a 3% chance to like Radio1, an 82% chance to like Radio2, a 34% chance to like Radio3 and a 92% chance to like Radio4. They know that 90% of the listeners are old.

Activity 2

Using the naive Bayes structure, the numerator above is given by

$$p(r_1 = \mathsf{like}|\mathsf{young})p(r_2 = \mathsf{dislike}|\mathsf{young})p(r_3 = \mathsf{like}|\mathsf{young})p(r_4 = \mathsf{dislike}|\mathsf{young})p(\mathsf{young}) \tag{10.1.5}$$

Plugging in the values we obtain

$$0.95 \times 0.95 \times 0.02 \times 0.8 \times 0.1 = 0.0014$$

The denominator is given by this value plus the corresponding term evaluated assuming the customer is old,

$$0.03 \times 0.18 \times 0.34 \times 0.08 \times 0.9 = 1.3219 \times 10^{-4}$$

Which gives

$$p(\mathsf{young}|r_1 = \mathsf{like}, r_2 = \mathsf{dislike}, r_3 = \mathsf{like}, r_4 = \mathsf{dislike}) = \frac{0.0014}{0.0014 + 1.3219 \times 10^{-4}} = 0.9161 \tag{10.1.6}$$

Naive Bayes

$$p(\mathbf{x}, c) = p(c) \prod_{i=1}^{D} p(x_i|c)$$

$$p(c|\mathbf{x}^*) = \frac{p(\mathbf{x}^*|c)p(c)}{p(\mathbf{x}^*)} = \frac{p(\mathbf{x}^*|c)p(c)}{\sum_c p(\mathbf{x}^*|c)p(c)}$$

$$\boxed{p(x) = \sum_y p(x, y) \qquad p(x|y) \equiv \frac{p(x, y)}{p(y)}}$$

Naive Bayes

Class-conditional probability

Prior probability

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

Posterior probability

The evidence

Naive Bayes

Class-conditional probability

Prior probability

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

Posterior probability

The evidence
Constant (same for all classes),
can be ignored

Naive Bayes

Prior probability

Can be computed from training data (fraction of records that belong to each class)

Class-conditional probability

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

Posterior probability

The evidence

Naive Bayes

Class-conditional probability

Method: Naive Bayes

Prior probability

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

Posterior probability

The evidence

Naive Bayes

- Mind that **X** is a vector

$$\mathbf{X} = \{X_1, \ldots, X_n\}$$

- Class-conditional probability

$$P(\mathbf{X}|Y) = P(X_1, \ldots, X_n|Y)$$

- "Naive" assumption: attributes are independent

$$P(\mathbf{X}|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

Naive Bayes

$$p(\mathbf{x}, c) = p(c) \prod_{i=1}^{D} p(x_i|c)$$

$$p(c|\mathbf{x}^*) = \frac{p(\mathbf{x}^*|c)p(c)}{p(\mathbf{x}^*)} = \frac{p(\mathbf{x}^*|c)p(c)}{\sum_c p(\mathbf{x}^*|c)p(c)}$$

$$\boxed{p(x) = \sum_y p(x, y) \qquad p(x|y) \equiv \frac{p(x, y)}{p(y)}}$$

Naive Bayes

Categorical attributes

- The fraction of training instances in class Y that have a particular attribute value x_i

$$P(X_i = x_i | Y = y) = \frac{n_c}{n}$$

n_c: number of training instances where X_i=x_i and Y=y

n: number of training instances where Y=y

Naive Bayes

The fraction of training instances in class Y that have a particular attribute value X_i

P(Status=Married|No)=?

P(Refund=Yes|Yes)=?

1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naive Bayes

- Can anything go wrong?

$$P(Y|\mathbf{X}) \propto P(Y) \prod_{i=1}^{n} P(X_i|Y)$$

What if this probability is zero?

If one of the conditional probabilities is zero,
then the entire expression becomes zero!

Naive Bayes

- Original

$$P(X_i = x_i | Y = y) = \frac{n_c}{n}$$

number of training instances where X_i=x_i and Y=y

number of training instances where Y=y

- Laplace smoothing

$$P(X_i = x_i | Y = y) = \frac{n_c + 1}{n + c}$$

c is the number of classes

Naive Bayes

- To highlight:
  - We consider the optimal Bayes classifier, which needs the true distributions
  - We approach it via naive Bayes
    - We assume naively the independence of the class-conditional attributes
    - We estimate $P(Xi|c), P(c)$
    - We ~assume very good data so no need to smooth, or we do smoothing