Esercitazione 7 24/11/2023

Apprendimento Bayesiano

Nell’ambito Bayesiano si cambia l’approccio avendo la valutazione di ipotesi in base alla loro probabilità. Si studia la probabilità rispetto ai dati e rispetto alle conoscenze pregresse. Non troviamo un’ipotesi che combacia ma che è probabile.

What is P?

- Frequentist vs Bayesian
  - An age-old debate, seemingly without an end in sight.
  - Both these point of view approach the same problem in different ways, which is why there is so much talk about which is better.

Frequentist vs Bayesian Interpretation

Long-Term Frequency	Degree of belief
Probability as the limit of the relative frequency of an event occurring in an infinite number of trials.	**Probability as a measure of belief** or certainty about an event, incorporating both prior knowledge and new evidence.

Probability as Long-Term Frequency

Key Points:

- **Objective Nature:** Focuses on the long-run frequency of events.
- **Fixed Parameters:** Assumes that parameters, such as the probability of an event, are fixed and not subject to uncertainty.
- **No Prior Beliefs:** Does not incorporate prior beliefs or subjective information.

Example: Frequentist perspective

θ is a constant and goal is to find the θ that would maximize the likelihood

> "I don't know what the mean female height is. However, I know that its value is fixed (not a random one). Therefore, I cannot assign probabilities to the mean being equal to a certain value, or being less than/greater than some other value. The most I can do is collect data from a sample of the population and estimate its mean as the value which is most consistent with the data.

P(Data|θ),

Probability as Degree of Belief

Key Points:
•**Subjective Nature:** Allows for the incorporation of subjective beliefs and prior knowledge.
• **Parameters & updating:** Parameters governed by probability distributions, updating as new evidence is acquired.
•**Prior Information:** Utilizes prior beliefs to update probabilities in light of new data.

- Remark: now θ is a variable, and the assumptions include a prior distribution of the hypotheses P(θ), and a likelihood of data P(Data|θ).

Example: Degree of belief

> "I agree that the mean is a fixed and unknown value, but I see no problem in representing the uncertainty probabilistically. I will do so by defining a probability distribution over the possible values of the mean and use sample data to update the distribution.

$$P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{P(Data)}$$

Focus: Updating probability

"I agree that the mean is a fixed and unknown value, but I see no problem in representing the uncertainty probabilistically. I will do so by defining a probability distribution over the possible values of the mean and use sample data to update the distribution.

- The statement that the sample data will be used to update the distribution is referring to Bayesian updating:
- The new data will make the probability narrower around the parameters true value through Bayes' theorem.

Example:
Before observing (e.g., the outcome of a coin toss), a Bayesian might assign a prior probability based on prior knowledge or beliefs.

Then (e.g., After the toss), prior is updated with the observed data.

Bayesian perspective: take home message

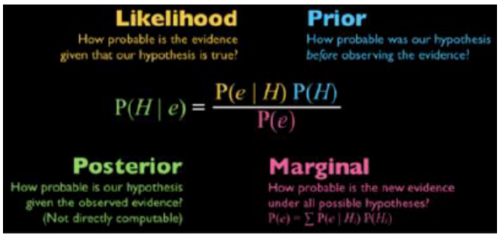
- Assume probabilities for both data and hypotheses (parameters)
  - For Bayesians, θ is a variable, and the assumptions include a prior distribution of the hypotheses P(θ), and a likelihood of data P(Data|θ).
    - Main issue: the subjectivity of the prior; different priors may arrive at different posteriors and conclusions.
  - After observation, this prior can be updated with new information (observed data).

Teorema di Bayes

- **Bayesians** estimate a full posterior distribution of the parameters using the Bayes' formula:

L'ipotesi è il mio modello, la mia supposizione.

Bayes formula & ML terminology



machine learning is interested in the *best hypothesis* h from some space H, given observed training data D

best hypothesis $\approx$ *most probable hypothesis*

Bayes Theorem provides a **direct method of calculating the probability of such a hypothesis** based on its prior probability, the probabilites of observing various data given the hypothesis, and the observed data itself

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

Consider parameter as our hypothesis, e.g.,

Ho:playtennis (yes)
H1:playtennis (no)
h in (0,1)

Prior p(h)

- How probable was our hypothesis, before observing the evidence?

è il grado di fiducia rispetto all'ipotesi. Prima di tutto si fissa la prior, dopo si fa l'osservazione.

- $P(h)$ *prior probability of* h, reflects any background knowledge about the chance that h is correct

Likelihood P(e | h)

How probable is the evidence given that our hypothesis is true?

- $P(D|h)$ probability of observing D given a world in which h holds

La verosimiglianza.

Evidence probability P(e)

- How probable is the new evidence under all possible hypotheses?
- REMARK: P(D) = E [P(D | H) P(H)]

- $P(D)$ *prior probability of* D, probability that D will be observed

La probabilità di evidenza.

Remark: evidence as the marignal computation over all possible values of the hypothesis

- The marginal probability of the evidence *P*(*e*) can be calculated by summing or integrating the joint probability of the hypothesis *H* and the evidence *E* over all possible values of *H*:

$P(B) = \sum_i P(B, A_i)$

In continuous cases, this is expressed as an integral:

$P(B) = \int P(B, A)\, dA$

This step is crucial for obtaining the posterior probability in Bayes' Theorem. The full Bayesian inference formula is:

$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

Posterior p(h|d)

- How probable is our hypothesis given the observed evidence?

REMARK : Not directly computable

- $P(h|D)$ ***posterior probability of*** h, reflects confidence that h holds after D has been observed

---

Bayes formula

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(h)$ = prior probability of hypothesis h
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

You might also wonder how to apply this formula practically
→ That's where **Maximum A Posteriori (MAP)** estimation comes into play.

L'ipotesi migliore è quella che rende massima la probabilità a posteriori.

MAP: Learning MAP hypotheses

in many learning scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis $h \in H$ given the observed training data D

$$\begin{aligned} h_{MAP} &= \underset{h \in H}{argmax}\ P(h|D) \\ &= \underset{h \in H}{argmax}\ \frac{P(D|h)P(h)}{P(D)} \\ &= \underset{h \in H}{argmax}\ P(D|h)P(h) \end{aligned}$$

note that $P(D)$ can be dropped, because it is a constant independent of h

Essendo il denominatore sempre uguale per ogni ipotesi, posso anche non considerarlo mentre ricerco l'ipotesi che rende la probabilità a posteriori massima.

Esempio

Example:
Suppose we want to know the probability of a student passing an exam (A) given that they attended a study group (B).

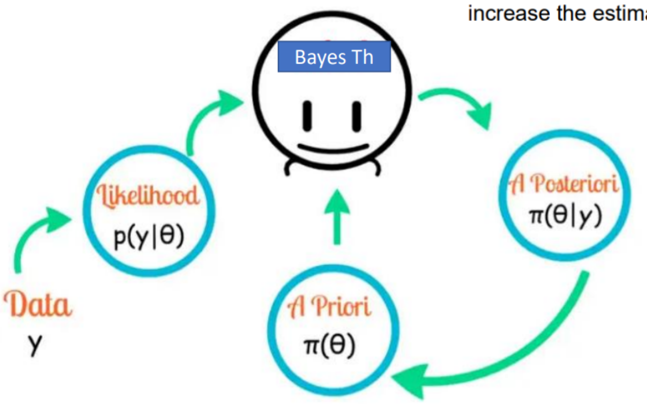
- **Given Information:**
  - $P(\text{pass}) = 0.70$ (Prior probability of passing the exam without considering the study group)
  - $P(\text{study group}|\text{pass}) = 0.90$ (Likelihood of attending a study group if the student passes)
  - $P(\text{study group}) = 0.60$ (Probability of attending a study group, irrespective of passing or failing)

- **Using Bayes' Theorem:**
  $P(\text{pass}|\text{study group}) = \frac{P(\text{study group}|\text{pass}) \cdot P(\text{pass})}{P(\text{study group})}$
- **Calculations:**
  $P(\text{pass}|\text{study group}) = \frac{(0.90) \cdot (0.70)}{0.60}$
- **Result:**
  $P(\text{pass}|\text{study group}) \approx \frac{0.63}{0.60} \approx 1.05$
- **Interpretation:**
  The updated probability of a student passing the exam, given that they attended a study group, is approximately 1.05. This suggests an increased likelihood of passing after considering the evidence from attending the study group.

qua forse i dati sono sbagliati

Focus: Probability updating again!

- each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct



$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

REMARK: **log** to simplify !

- Since logarithmic functions are monotonic, we can rewrite the above equation in the log space and decompose it into 2 parts: maximizing the likelihood and maximizing the prior distribution:

$$
\begin{aligned}
h_{MAP} &= \underset{h \in H}{argmax}\; P(h|D) \\
&= \underset{h \in H}{argmax}\; \frac{P(D|h)P(h)}{P(D)} \\
&= \underset{h \in H}{argmax}\; P(D|h)P(h)
\end{aligned}
$$

$$
\begin{aligned}
h_{MAP} &= \mathrm{argmax}_\theta(\log P(Data|h) + \log P(h)) \\
&= \mathrm{argmax}_\theta(L(h) + \log P(h))
\end{aligned}
$$

Esempio tennis

- Today's weather forecast: play yes or not?

Input:
Temperature = Hot (H)
Wind = Weak (W)

Should I play tennis?

Prior

Play tennis	P(Play tennis)
Yes	0.3
No	0.7

Likelihood

Temperature	Wind	P(T, W |Tennis = Yes)
Hot	Strong	0.15
Hot	Weak	0.4
Cold	Strong	0.1
Cold	Weak	0.35

Temperature	Wind	P(T, W |Tennis = No)
Hot	Strong	0.4
Hot	Weak	0.1
Cold	Strong	0.3
Cold	Weak	0.2

argmax_y P(H, W | play?) P (play?)

P(H, W | Yes) P(Yes) = 0.4 $\times$ 0.3
= 0.12

P(H, W | No) P(No) = 0.1 $\times$ 0.7
= 0.07

MAP prediction = Yes

The complexity of learning «play tennis»

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Outlook: S(unny), O(vercast), R(ainy)

Temperature: H(ot), M(edium), C(ool)

Humidity: H(igh), N(ormal), L(ow)

Wind: S(trong), W(eak)

We need to ESTIMATE

1. The prior $P(\text{Play?})$
2. The likelihoods $P(\text{x} \mid \text{Play?})$

Outlook: S(unny), O(vercast), R(ainy)
Temperature: H(ot), M(edium), C(ool)
Humidity: H(igh), N(ormal), L(ow)
Wind: S(trong), W(eak)

	O	T	H	W	Play?
1	S	H	H	W	-
2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
8	S	M	H	W	-
9	S	C	N	W	+
10	R	M	N	W	+
11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-
	3	3	3	2	

Values for this feature

Prior P(play?)

- A single number (Why only one?)

Likelihood P(**X** | Play?)

- There are 4 features
- For each value of Play? (+/-), we need a value for each possible assignment: P(O, T, H, W | Play?)

- $(3 \cdot 3 \cdot 3 \cdot 2 \; - 1)$ parameters in each case

  One for each assignment

Ne basta uno perché calcolo solo al probabilità di giocare, quella di non giocare la ottengo facendo poi 1 - prob di giocare.

In general

Prior P(Y)

- If there are k labels, then k – 1 parameters (why not k?)

Likelihood P(**X** | Y)

- If there are d Boolean features:
  - We need a value for each possible $P(x_1, x_2, \cdots, x_d \mid y)$ for each y
  - $k(2^d - 1)$ parameters

Need a lot of data to estimate these many numbers!

How can we deal with this?

Answer: Make independence assumptions

Next steps

- Naive Bayes
- Learning continuos features with Bayes (Gaussian Naive Bayes)