Probabilistic Methods

CMSC 422

Slides adapted from Prof. CARPUAT

Today's topics

- Bayes rule review
- A probabilistic view of machine learning
  - Joint Distributions
  - Bayes optimal classifier
- Statistical Estimation
  - Maximum likelihood estimates
  - Derive relative frequency as the solution to a constrained optimization problem

Bayes Rule

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

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**

we call P(A) the "prior"

and P(A|B) the "posterior"

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning*…

Exercise: Applying Bayes Rule

- Consider the 2 random variables

  A = You have the flu

  B = You just coughed

- Assume

  $P(A) = 0.05$

  $P(B|A) = 0.8$

  $P(B|\text{not } A) = 0.2$

- What is $P(A|B)$?

Answer

- Via logic
  - Assume 100 students – 5 have the flu. 80% (4) of the students who have the flu cough; 20% (19) of the students who don't have the flu cough; So the chance that you have the flu is 4/23
- Via Bayes Rule:
  - $P(A|B)P(B)=P(B|A)P(A)$.
  - $P(B)=0.8*0.05+0.2*(1-0.05)=0.04+0.19=0.23$
  - $P(A|B)=0.8*0.05/0.23 =0.04/0.23=4/23$

Q: What does this have to do with machine learning ?

Using a Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Using a Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

- Given the joint distribution, we can find the probability of any logical expression E involving these variables

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using a Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Given the joint distribution, we can make inferences

- E.g., P(Male|Poor)?
- Or P(Wealth | Gender, Hours)?

Recall: Machine Learning as Function Approximation

Problem setting

- Set of possible instances X
- Unknown target function $f: X \rightarrow Y$
- Set of function hypotheses $H = \{h \mid h: X \rightarrow Y\}$

Input

- Training examples $\{(x^{(1)}, y^{(1)}), \dots (x^{(N)}, y^{(N)})\}$ of unknown target function f

Output

- Hypothesis $h \in H$ that best approximates target function f

Recall: Formal Definition of Binary Classification (from CIML)

TASK: BINARY CLASSIFICATION

Given:

1. An input space $\mathcal{X}$
2. An unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$

Compute: A function f minimizing: $\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[f(\boldsymbol{x}) \neq y\big]$

The Bayes Optimal Classifier

- Assume we know the data generating distribution $\mathcal{D}$
- We define the **Bayes Optimal classifier** as

$$f^{(\text{BO})}(\hat{x}) = \arg\max_{\hat{y} \in \mathcal{Y}} \mathcal{D}(\hat{x}, \hat{y})$$

- **Theorem:** Of all possible classifiers, the Bayes Optimal classifier achieves the smallest zero/one loss
- **Bayes error rate**
  - Defined as the error rate of the Bayes optimal classifier
  - Best error rate we can ever hope to achieve under zero/one loss

If we had access to $\mathcal{D}$, Finding an optimal classifier would be trivial! we don't have access to $\mathcal{D}$. So let's try to estimate it instead!

What does “training” mean in probabilistic settings?

- Training = estimating $\mathcal{D}$ from a finite training set
  - We typically assume that $\mathcal{D}$ comes from a specific family of probability distributions

    e.g., Bernouilli, Gaussian, etc
  - Learning means inferring parameters of that distributions

    e.g., mean and covariance of the Gaussian

Training assumption: training examples are iid

- **Independently and Identically distributed**
  - i.e. as we draw a sequence of examples from $\mathcal{D}$, the n-th draw is independent from the previous n-1 sample
- This assumption is usually false!
  - But sufficiently close to true to be useful

How can we estimate the joint probability distribution from data?

What are the challenges?