A Probabilistic View of Machine Learning (1/2)

CMSC 422

MARINE CARPUAT
marine@cs.umd.edu

Some slides based on material by Tom Mitchell

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

we call P(A) the "prior"

and P(A|B) the "posterior"

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**

…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 $P(A|B)$?

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: 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: $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[f(\boldsymbol{x}) \neq y\right]$

The Bayes Optimal Classifier

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

$$f^{(\mathrm{BO})}(\hat{x}) = \arg\max \mathcal{D}(\hat{x}, \hat{y})$$

- **Th**
  cl
- **B**
  –
  – 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

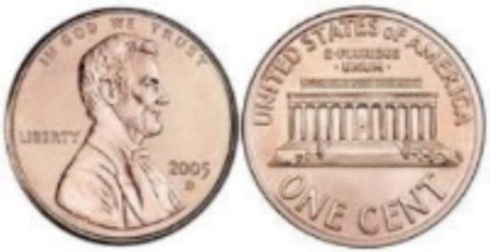
Training assumption: training examples are iid

- **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?

- Challenge: sparse and incomplete observations
- One approach: maximum likelihood estimation
  - Finds the parameters that maximize the probability of the data

Maximum Likelihood Estimates



X=1 X=0

P(X=1) = θ

P(X=0) = 1-θ

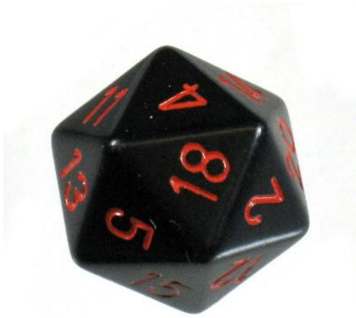
(Bernoulli)

Given a data set D of iid flips, which contains α_1 ones and α_0 zeros

$$P_\theta(D) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}_{MLE} = argmax_\theta \; P_\theta(D) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Maximum Likelihood Estimates



K sided die
$\forall\, k, P(X = k) = \theta_k$

(Categorical Distribution)

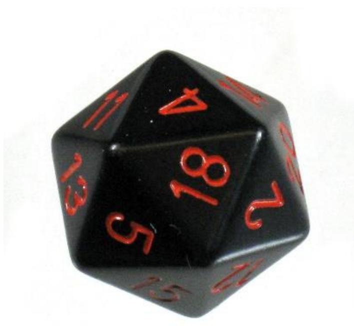
Given a data set D of iid rolls, which contains x_k outcomes for each k

$$P_\theta(D) = \prod_{k=1}^{K} {\theta_k}^{x_k}$$

$$\begin{aligned}\hat{\theta}_{MLE} &= argmax_\theta \; P_\theta(D) \\ &= argmax_\theta \; \log \; P_\theta(D) \\ &= argmax_\theta \sum_{k=1}^{K} x_k \log(\theta_k)\end{aligned}$$

Problem:
This objective lacks constraints!

Maximum Likelihood Estimates



K sided die

$\forall\, k, P(X = k) = \theta_k$

A constrained optimization problem

$$\hat{\theta}_{MLE} = argmax_{\theta} \sum_{k=1}^{K} x_k \log(\theta_k)$$

$$with \quad \sum_{k=1}^{K} \theta_k = 1$$

We can solve this using **Lagrange multipliers** (on board)

$$\hat{\theta}_k = \frac{x_k}{\sum_i x_i}$$

What you should know

- Bayes rule
- A probabilistic view of machine learning
  - If we know the data generating distribution, we can define the Bayes optimal classifier
  - Under iid assumption
- How to estimate a probability distribution from data?
  - Maximum likelihood estimates
  - for Bernoulli and Categorical distributions