INTRODUCTION TO MACHINE LEARNING COMPSCI 4ML3

Lecture 9

Hassan Ashtiani

REVIEW: BAYES RULE, CHAIN RULE

$P(X = a)$ ↔ $P(X)$

- Joint distribution: $P(X, Y)$
- Sum rule: $P(X) = \sum_{Y} P(X, Y)$
- Conditional distribution: $P(X|Y) = \frac{P(X,Y)}{P(Y)}$
- Bayes rule: $P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$
- Chain Rule: $P(X_1, X_2, \ldots, X_k) =$
  $P(X_1)P(X_2|X_1)P(X_3|X_2,X_1), \ldots, P(X_k|X_1,X_2,\ldots X_{k-1})$
- $P(X_1, X_2, \ldots, X_k | Y) = P(X_1|Y)P(X_2|X_1,Y) \ldots$
- $P(X|Y), P(Y|X) \rightarrow P(X,Y) = ?$

REVIEW: INDEPENDENCE

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

- X AND Y ARE INDEPENDENT IF $P(X,Y)$ $= P(X)P(Y)$

ASSUME $X_1, \dots, X_k$ ARE INDEPENDENT GIVEN Y

- $P(X_1, X_2, \dots, X_k | Y) =$

  $P(X_1|Y)\,P(X_2|Y) \cdots P(X_k|Y)$

-

STATISTICAL APPROACH TO ML

- Our goal is to do well on new/unseen (test) data
- We were mostly minimizing the training error
  - Directly/systematically optimizing the test error?
- There is uncertainly about the unseen data
  - We cannot be 100% sure about the performance of <u>any method</u> on the test data
  - A method that works well on test set <u>**most of the time**</u>?

I.I.D ASSUMPTION

- Assume there is an underlying (unknown) distribution D
- Assume that each of the training and test instances are sampled independently from D
- We say train and test sets are i.i.d. (independent and identically distributed) samples generated from D

I.I.D Assumption

- Why are these assumptions necessary?
  - Same distribution for all samples ("identically")
  - Independent samples ("Independently")
  - Same distribution for train and test
  - The distribution is unknown

PARAMETER ESTIMATION

- Assume that the distribution comes from some known family
  - Bernoulli, Gaussian, …
- Use the training set to estimate the value of the unknown distribution parameters
- Useful in both unsupervised and supervised learning

BIASED COIN EXAMPLE (UNSUPERVISED)

- FLIPPING A COIN
  - OUTCOME IS HEAD (0) OR TAIL (1), SO $x \in \{0,1\}$
- $P(x = 0) = \alpha$, $P(x = 1) = 1 - \alpha$
  - x IS A BERNOULLI RANDOM VARIABLE
- BIAS (α) IS UNKNOWN (THE PARAMETER)
- GIVEN AN I.I.D SAMPLE, **ESTIMATE α**
  - $X = (x^1, x^2, x^3, \ldots, x^n)$
  - E.G., $n = 10$, $X = (0,0,1,1,0,1,0,1,0,0)$

$\hat{\alpha} = \frac{6}{10}$

ESTIMATING THE BIAS OF THE COIN

- LET n_0 =#HEADS, n_1 = #TAILS (SO $n_0 + n_1 = n$)
- IS $\hat{\alpha} = \frac{n_0}{n_0+n_1}$ A GOOD ESTIMATE?
- IS THERE A RATIONAL BEHIND THIS ESTIMATE?

MAXIMUM LIKELIHOOD ESTIMATE (MLE)

GIVEN THE TRAINING SET $X = (x^1, x^2, \ldots, x^n)$, ESTIMATE α.

- MLE MAXIMIZES THE PROBABILITY OF THE OBSERVATIONS GIVEN THE PARAMETERS
  - $\alpha^{ML} = \underset{\alpha}{argmax}\, P(X|\alpha)$

likelihood

- EQUIVALENTLY (WHY?)
  - $\alpha^{ML} = \underset{\alpha}{argmin} - \log(\sum_i P(x^i|\alpha))$

$- \sum \log(P(x^i|\alpha))$

NEGATIVE-LOG-LIKELIHOOD

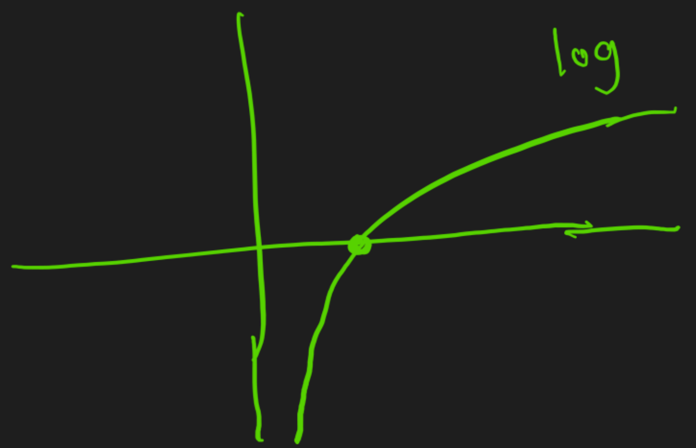
$$\alpha^{ML} = \arg\max_{\alpha} p(x|\alpha) = \arg\min_{\alpha} -p(x|\alpha)$$

$$= \arg\min_{\alpha} -\log p(x|\alpha)$$

$$= \arg\min_{\alpha} -\log \prod_{i=1}^{n} p(x^i|\alpha)$$

$$= \arg\min_{\alpha} -\sum_{i=1}^{n} \log p(x^i|\alpha) \checkmark$$

monotone



MAXIMUM LIKELIHOOD FOR COINS

$$\text{Likelihood} = p(X|\alpha) = \prod_{i=1}^{n} p(x^i|\alpha) =$$

$$= \left(\prod_{i:\, x^i=0} \underbrace{p(x^i|\alpha)}_{\alpha} \right) \left(\prod_{i:\, x^i=1} \underbrace{p(x^i|\alpha)}_{1-\alpha} \right)$$

$$= \alpha^{n_0} (1-\alpha)^{n_1} = f(\alpha)$$

$$0 = \frac{\partial f}{\partial \alpha} = n_0 \alpha^{n_0-1} (1-\alpha)^{n_1} - \alpha^{n_0} n_1 (1-\alpha)^{n_1-1}$$

$$\Longrightarrow n_0(1-\alpha) - n_1 \alpha = 0 \rightarrow \boxed{\alpha^{ML} = \frac{n_0}{n_0+n_1}}$$

MAXIMUM A POSTERIORI ESTIMATE

- MAXIMIZES THE PROBABILITY OF THE PARAMETERS GIVEN THE OBSERVATIONS
    - $\boldsymbol{\alpha}^{MAP} = \underset{\boldsymbol{\alpha}}{\boldsymbol{argmax}}\, \boldsymbol{P}\,(\boldsymbol{\alpha}|\boldsymbol{X})$
- $\alpha^{MAP} = \underset{\alpha}{argmin}\left(-\text{LOG}(P(\alpha)) - \sum_{i=1}^{n} \text{LOG}\, \boldsymbol{P}(x^i|\alpha)\right)$

PRIOR VS POSTERIOR DISTRIBUTIONS

- $\mathrm{P}(\alpha)$ CAPTURES THE **PRIOR** DISTRIBUTION
- $\mathrm{P}(\alpha|X)$ CAPTURES THE **POSTERIOR** DISTRIBUTION
- IN OTHER WORDS,
  - WE START BY A PRIOR BELIEF ABOUT VALUE OF α
  - OUR BELIEF IS UPDATED AFTER SEEING SOME REAL DATA
  - THIS IS A **BAYESIAN** APPROACH

MAP FOR COINS – UNIFORM PRIOR

MAP FOR COINS – NONUNIFORM PRIOR