9/4/19 Learning Machine Learning

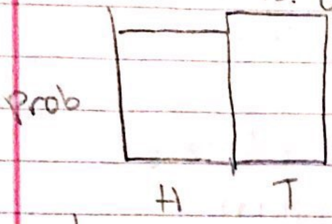
ex) heads v. tails

Probability of heads measured: $P_{meas}(H) = \lim_{T \to \infty} \#H / \#T$

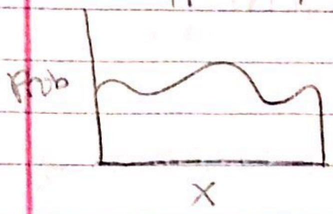
T: # of tails

→ frequentist approach

· vs. degree of belief (Bayesian)



$\boxed{\sum_i^N P_i = 1}$ (discrete)



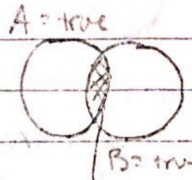
$\boxed{\int P(x)dx = 1}$ (continuous)

↳ probability distribution

sum rule

· A, B = true/false

· $\underbrace{P(A+B)}_{A \vee B \ (A \cup B)} = P(A) + P(B) - \underbrace{P(A \& B)}_{A \cap B}$



↳ minus overlap: P(A+B)

· $P(A\&B) = P(A \cap B) = P(A)\underbrace{P(B|A)}_{\text{prob. of B assuming A (conditional)}} = P(B)P(A|B)$ (product rule)

↓

· Bayes' thm.: $P(A|B) = P(A)P(B|A)/P(A)$

$P(\alpha|X) = P(\alpha)P(X|\alpha)/P(X)$

posterior, event, ↳ data, ↳ prior, ↳ likelihood, ↳ normalization/evidence

· likelihood: how consistent is the evidence w/ possibility α

· prior: prior belief in possibility (common sense)

Types of ML applications

- classification: given data, classify it
- regression: given data, learn function (model) (fitting)
- density estimation (unsupervised): given data, fit/cluster data without labels

ex) $P(\alpha|X)$ hypothesis testing

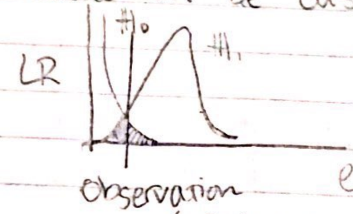
↳ classify which possibility is most likely

- Neumann-Pearson lemma: universally most powerful test statistic is the likelihood ratio: $LR = P(H_0|X)/P(H_1|X)$

H_0: null hypothesis

H_1: any other possibility

→ how should LR be distributed in an imaginary ensemble of events?



↳ can make a statement on which H you prefer and how strongly you prefer it

→ see slide 2 for graph on hypothesis testing

p values { • β: probability of observing data if H_1 is true
• α: " " H_0 "

- type 1 error: α threshold set in advance (i.e. $p = 0.05$) (chance of favoring H_1 if H_0 is true)
- type 2 error: favoring H_0 when H_1 is true

→ frequentist (in comic) only looks @ likelihood & ignores prior

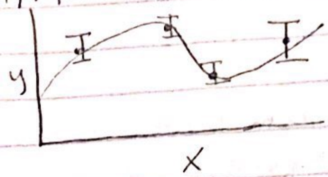
SN = supernova

- $P(SN|\text{"yes"}) = P(SN)P(\text{"yes"}|SN) / P(\overline{SN})P(\text{"yes"}|\overline{SN}) + P(SN)P(\text{"yes"}|SN)$

Bayesian { frequentist { • $P(\text{"yes"}|SN) = 35/36$
• $P(\text{"yes"}|\overline{SN}) = 1/36$
• $P(SN) \sim 10^{-9}$
• $P(\overline{SN}) \sim 1 - 10^{-9}$ }

LML 9/4/19

ex) fake data fitting



- assume Gaussian errors
→ likelihood? $\mathcal{L}_{DATA}(\vec{\alpha}) = \prod_i^4 P(x_i|\vec{\alpha}) = \prod_i^4 1/\sqrt{2\pi\sigma_i^2}\, e^{-(x_i - \hat{y}(x_i))^2/2\sigma_i^2}$
↳ parameters ↳ variance

- x_i: data points
- $\hat{y}(x_i)$: predicted (model) value

→ log likelihood: $\log\mathcal{L} = \sum_i (x_i - \hat{y}(x_i))^2/2\sigma_i^2 + c^{te}$
↳ chi-squared distribution: used to evaluate goodness of fit

→ maximize likelihood = minimize chi-squared
↳ can minimize χ^2 further by making model more flexible (more parameters/dof)
⇒ more complex model → more volume of possible parameter values $P(x)$

- minimize expectation value (avg.) $\mathbb{E}[(\underbrace{\hat{y}(x_i)}_{\text{model}} - \underbrace{y_i}_{\text{data label}})^2]$ (RSS, MSE) (supervised learning)
↳ can rearrange to get "variance" + "bias" terms
- bias: systematic shortcomings that make model incapable of fitting data
- variance: model fits data (too) well, but high variance leads to low generalizability
↳ more parameters

ex) Slide 3: Decision problem (classifier)
- training a classifier to classify orange/blue
- Bayes' decision boundary = optimal boundary

Slide 4: math (simple least squares model)
- linear least squares model: $\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} x_j \hat{\beta}_j$
↳ soln.: $\beta = (X^T X)^{-1} X^T y$ (optimal parameters)

Scanned with CamScanner

CML

ex) Con't

- K-nearest neighbors (slide 6)
  - Clustering: choose nearest neighbors & average
  - Slide 7: boundary for k=15 (hyperparameter)
    - ↳ low variance, high bias for high k (underfitting)
    - ↳ vice versa for low k (overfitting)
- Slide 8: algorithm performance