PROBABILITY

OFTEN THE GOAL OF MACHINE LEARNING IS TO DETERMINE THE PROBABILITY OF AN EVENT.

EXAMPLES

- Pr(YEAR THAT POLAR ICE CAP MELTS ≤ 2030)
- Pr(A NEW EMAIL IS SPAM)
- Pr(A PERSON IS AT RISK FOR A DISEASE)

□

STATISTICS

THERE ARE TWO MAIN PARADIGMS IN STATISTICS:

1. FREQUENTISTS
2. BAYESIAN

1. FREQUENTISTS

- PROBABILITIES ARE LONG RUN FREQUENCIES.
  EXAMPLE: FLIP A COIN MILLION TIMES TO DETERMINE IF ITS FAIR

- PROBABILITY OF AN EVENT IS DEFINED AS THE FREQUENCY THAT SPECIFIC EVENT OCCURRED IN A LONG LIST OF REPEATED TRIALS.

2. BAYESIAN

- PROBABILITIES QUANTIFY OUR UNCERTAINTY IN EVENTS DESIGNED TO GET THE CLOSEST TO THE TRUTH GIVEN A SPECIFIC SET OF DATA.

- WE HAVE TO WORK WITH THE GIVEN SET OF DATA POINTS. WE DO NOT HAVE THE LUXURY OF OBSERVING MULTIPLE TRIALS. WE THEREFORE REQUIRE THE ABILITY TO GENERALIZE FROM A SMALL NUMBER OF EVENTS.

BAYESIAN STATISTICS

$\mathcal{D}$ = SET OF DATA POINTS GENERATED FROM SOME DISTRIBUTION PARAMETERIZED BY θ.

WE WANT TO ESTIMATE θ

THAT IS, WE WANT TO FIND MOST LIKELY VALUE OF θ GIVEN $\mathcal{D}$.

THIS IS $\underset{\theta}{\text{ARG MAX}}\ P(\theta|\mathcal{D})$

AS PER BAYES' RULE

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \quad ; \quad P(\mathcal{D}) > 0$$

THE TERMS HAVE THE FOLLOWING MEANING.

- $P(\theta|\mathcal{D})$ = POSTERIOR DISTRIBUTION OF θ GIVEN $\mathcal{D}$
- $P(\mathcal{D}|\theta)$ = LIKELIHOOD OF OBSERVATIONS, GIVEN THE MODEL
- $P(\theta)$ = PRIOR DISTRIBUTION OR BELIEF
- $P(\mathcal{D})$ = EVIDENCE, OR MARGINAL LIKELIHOOD OF $\mathcal{D}$.
  IT IS A CONSTANT WITH RESPECT TO θ.

- $P(\theta|\mathcal{D})$ = POSTERIOR DISTRIBUTION OF θ GIVEN $\mathcal{D}$ AND CALCULATING (OR APPROXIMATING) IS THE MAIN GOAL OF BAYESIAN INFERENCE

IT EXPRESSES UNCERTAINTY ABOUT θ, AFTER WE HAVE SEEN ("LEARNED FROM") THE DATA. THERE IS STILL UNCERTAINTY BECAUSE WHILE THE DATA TOLD US SOMETHING ABOUT θ, WE STILL DO NOT KNOW EVERYTHING ABOUT IT.

- $P(\mathcal{D}|\theta)$ = LIKELIHOOD. IT IS A FUNCTION OF θ.

THIS IS A WAY OF DESCRIBING THE PROBABILITY OF THE DATA AS A FUNCTION OF THESE UNKNOWN PARAMETERS.

- $P(\theta)$ = PRIOR DISTRIBUTION. IT DESCRIBES OUR BELIEF ABOUT THE QUANTITY OF INTEREST BEFORE WE SEE DATA. OFTEN TIMES WE USE WHAT ARE CALLED CONJUGATE PRIORS TO THE LIKELIHOOD TO DESCRIBE OUR A PRIORI BELIEFS.

- $P(\mathcal{D})$ = MARGINAL LIKELIHOOD OF $\mathcal{D}$, AND IT IS CONSTANT WITH RESPECT TO θ.

THEREFORE WE WRITE $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta)$

↑ PROPORTIONALITY SYMBOL

$$P(\mathcal{D}) = \int_{\theta} P(\mathcal{D}|\theta)P(\theta)\,d\theta$$

THIS IS CALLED THE MARGINAL LIKELIHOOD, BECAUSE WE ARE MARGINALISING (OR AVERAGING) OVER $\theta \in \Theta$

LOWER-CASE θ ↑ UPPER-CASE Θ