INF367A: Probabilistic machine learning

Lecture 5: Bayesian modeling I

Pekka Parviainen

University of Bergen

4.2.2020

Outline

Probability theory is nothing but common sense reduced to calculation. - Pierre Laplace, 1814

Probability interpretations

Does it make sense to say $P(A) = 0.7$ or 0.2 for the following A?

- $A = \{$ A patient recovers from cancer $\}$
- $A = \{$ It will rain tomorrow $\}$
- $A = \{$ It rained this day last year $\}$
- $A = \{$ A coin comes up heads $\}$
- $A = \{$ There is life beyond earth $\}$
- $A = \{$ Finland will win European football championship in 2020 $\}$
- $A = \{$ Ålesund has more inhabitants than Molde $\}$

Uncertainty

- Aleatory uncertainty
  - Due to randomness
  - We are not able to obtain observations which can reduce this uncertainty
- Epistemic uncertainty
  - Due to lack of knowledge
  - We are able to obtain observations which can reduce this uncertainty
  - Two observers may have different epistemic uncertainty

Probability interpretations

Two commonly used interpretations:

- Frequentist, objective
  - Frequencies from repetitions of experiments (realizable or hypothetical)
  - Handles aleatory uncertainty
- Bayesian, subjective, degree of belief
  - Here A are propositions and $P(A)$ is a degree of belief in A being true.
  - We may say "I believe to the extent of $P(A)$ that A is true".
  - Contrast with the frequentist interpretation: there $P(A)$ is the proportion of times that A occurs to be true.
  - Handles both aleatory and epistemic uncertainty

In the earlier slide, which interpretation might be applied to $P(A)$?

Bayesian inference

- Interpret probability as a degree of belief
- Basic idea:
  - Start with your initial beliefs
  - Observe evidence
  - Update your beliefs based on evidence
- Uncertainty is represented by probability distributions

Bayes' theorem revisited

- We want the distribution of the parameters given the observed data:

$$P(\text{model} \mid \text{data})$$

- We can use the Bayes theorem:

$$P(\text{model} \mid \text{data}) = \frac{P(\text{data} \mid \text{model})P(\text{model})}{P(\text{data})}$$

- $P(\text{model} \mid \text{data})$: Posterior probability of parameters after observing data
- $P(\text{data} \mid \text{model})$: Likelihood
- $P(\text{model})$: Prior probability of parameters before observing data
- $P(\text{data})$: Normalizing constant

Prior distribution

- $P(\theta)$: Your belief about plausibility of θ before observing data
  - But contains all of your prior knowledge
- Subjective: Your prior may differ from mine
  - Different priors $\Rightarrow$ different posterior
  - The more data you have, the smaller the effect of the prior will be
- Why does it make sense to have a prior?
  - Incorporate prior knowledge
  - Regularization

Different types of priors

- Cromwell's rule: If $P(\theta) = 0$, then the posterior $P(\theta \mid D)$ is always zero (similarly, if $P(\theta) = 1$, then posterior is always 1)
- Uninformative/objective/reference prior
  - Uniform, as "wide" as possible.
  - Principle of indifference: all possibilities have an equal probability.
- Informative prior
  - Not uniform
  - Assumes that we have some prior knowledge
- Conjugate prior
  - Prior and posterior have the same type of distributions (given that likelihood is of certain type)
  - Simplifies the computations
  - More later ...

Likelihood

- $P(D\,|\,\theta)$ is the probability that the model generates the observed data D when using parameter θ
  - $L(\theta) \equiv P(D\,|\,\theta)$, with D held fixed, is called the *likelihood*
  - $f(y) \equiv P(D\,|\,\theta)$, with θ held fixed, is called the *observation model* or the *sampling distribution*

Maximum likelihood estimation

- "Standard" machine learning approach
- $\theta_{ML} = \arg\max_\theta P(D \,|\, \theta)$
- Commonly used
  - For example, linear regression, neural networks
- A point estimate, does not quantify our uncertainty about θ

Normalizing constant $P(D)$

- Also called *evidence* or *marginal likelihood*
- Discrete parameters: $P(D) = \sum_\theta P(D, \theta)$
- Continuous parameters: $P(D) = \int_\theta P(D, \theta) \mathrm{d}\theta$
- Challenge: Typically too complex to be computed

Posterior distribution

$$P(\theta \,|\, D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}$$

- Posterior is the result of Bayesian inference
- Tells of the uncertainty related to the value of θ after observing D

Bayes' theorem as an update rule

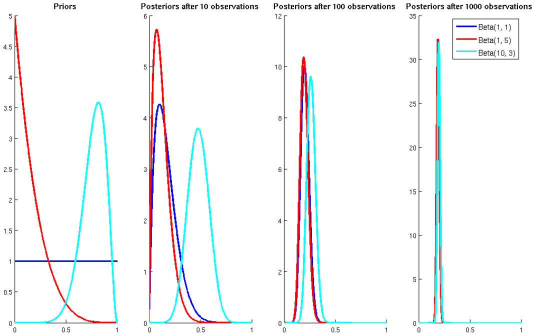
- Exchangeability: A sequence of random variables $X_1, X_2, X_3, \ldots$ is *exchangeable* if their joint distribution does not change when the positions in the sequence are changed
  - For example, independent and identically distributed variables
- For exchangeable variables it holds that

$$
\begin{aligned}
P(\theta \mid D_1, D_2) &= \frac{P(D_1, D_2 \mid \theta) P(\theta)}{P(D_1, D_2)} \\
&= \frac{P(D_1 \mid D_2, \theta) P(D_2 \mid \theta) P(\theta)}{P(D_1 \mid D_2) P(D_2)} \\
&= \frac{P(D_2 \mid D_1, \theta) P(D_1 \mid \theta) P(\theta)}{P(D_2 \mid D_1) P(D_1)} \\
&= \frac{P(D_2 \mid D_1, \theta)}{P(D_2 \mid D_1)} \cdot \frac{P(D_1 \mid \theta) P(\theta)}{P(D_1)} \\
&= \frac{P(D_2 \mid D_1, \theta)}{P(D_2 \mid D_1)} \cdot P(\theta \mid D_1) \\
&= \frac{P_1(D_2 \mid \theta) P_1(\theta)}{P_1(D_2)}
\end{aligned}
$$

Sequential Bayesian updating

- Several equivalent ways to compute the posterior $P(\theta \mid D_1, D_2)$
  - Use prior $P(\theta)$, observe $D = \{D_1, D_2\}$, and compute $P(\theta \mid D)$
  - Use prior $P(\theta)$, observe D_1 and compute $P(\theta \mid D_1)$. Then use prior $P(\theta \mid D_1)$, observe D_2, and compute $P(\theta \mid D_1, D_2)$
  - Use prior $P(\theta)$, observe D_2 and compute $P(\theta \mid D_2)$. Then use prior $P(\theta \mid D_2)$, observe D_1, and compute $P(\theta \mid D_1, D_2)$
- "Today's posterior is tomorrow's prior"
- Advantage: you can learn online and do not need to store data

Effect of the prior



- When we have little data, the choice of the prior has large effect
- The more data we have, the smaller the effect of the prior
  - The stronger the prior, the more data you need overdrive the prior

Example: Bernoulli model

- Jupyter notebook: Bernoulli_model.ipynb

Predictive distribution

- Prediction in standard machine learning:
  - Find a model θ given data D
  - Make predictions with θ
- Bayesian prediction:

$$\begin{aligned}
P(d_{new} \mid D) &= \int_{\theta\in\Theta} P(\theta, d \mid D)\mathrm{d}\theta \\
&= \int_{\theta\in\Theta} P(d_{new} \mid \theta, D)P(\theta \mid D)\mathrm{d}\theta \\
&= \int_{\theta\in\Theta} P(d_{new} \mid \theta)P(\theta \mid D)\mathrm{d}\theta
\end{aligned}$$

- Bayesian prediction uses predictions $P(d_{new} \mid \theta)$ from all the models θ, and weighs them by the posterior probability $P(\theta \mid D)$ of the models

Predictive distribution

- Often we cannot compute the predictive distribution analytically
- Solution: Monte Carlo approximation
  1. For $s = 1, \ldots S$:
     1.1 Sample parameter values from the posterior: $\theta_s \sim P(\theta \mid D)$
     1.2 Sample a data point given the parameter from the sampling distribution (likelihood): $x_{new} \sim P(x \mid \theta_s)$
- The predictive distribution is represented by the samples

Summarising the posterior

- Sometimes it is not convenient to present the results as a full posterior.
  - For example, if we have lots of parameters.
- We may be interested in only a handful of parameters or we will use the results for a particular task.
- Thus, it may be more convenient to summarize the posterior.

Point estimates

- Sometimes we want to collapse the posterior into a single point.
- Maximum a posteriori (MAP) estimate (the most likely value)

$$\theta_{MAP} = \arg\max_{\theta} P(D \mid \theta)P(\theta)$$

- MAP estimate with a uniform prior is equal to the ML estimate
- Downsides:
  - No uncertainty measure
  - May overfit
  - Mode may be an untypical point

Confidence intervals

- 95 % Bayesian confidence interval (or credible interval) is an interval where there is a 95% probability that the parameter is within the interval
  - Contrast to the frequentist approach where a confidence interval is a range where the statistic is 95% of the samples (assuming ML-estimate is correct)
- Not unique: may be centered with the median or the mean as a center point

Why the Bayesian way makes more sense than the frequentist way (my subjective view)

- Suppose we have a null hypothesis H_0 and an alternative hypothesis H_1. We want to know which one is true
- Frequentist way:
  - Compute p value: the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct.
- Bayesian way:
  - Compute posterior probability $P(H_0 \mid D)$: the probability that H_0 is true given the observed data
- Typically, Bayesians answer the questions that you would like to ask

Further readings

- Bishop 1.2.3, 2.1
- Hall: Bayesian inference