MLE&MAP &NAIVE BAYES

FEB. 7, 2019

KELLY SHI

OUTLINE

- Probability quick review
- Bayes Theorem
- MAP&MLE
- Naive Bayes Classifier

PROBABILITY REVIEW

- conditional probability: $\mathrm{P}[A \mid B] = \frac{P[A \cap B]}{P[B]}$
- total probability: $\mathrm{P}[A] = \sum_i \mathrm{P}[A \mid B_i] \cdot \mathrm{P}[B_i]$
- chain rule: $\mathrm{P}\left[\bigcap_{i=1}^{N} A_i\right] = \prod_{i=1}^{N} \mathrm{P}\left[A_i \mid \bigcap_{j=1}^{i-1} A_i\right]$
- A,B are independent if $\mathrm{P}[A \cap B] = \mathrm{P}[A] \cdot \mathrm{P}[B]$
- A,B are conditionally independent, given C, if

$$\mathrm{P}[A \cap B \mid C] = \mathrm{P}[A \mid C] \cdot \mathrm{P}[B \mid C]$$

BAYES THEOREM

- Using chain rule: $P[A \cap B] = P[A \mid B]P[B] = P[B \mid A]P[A]$
- Rearrage we get: $P[A \mid B] = \dfrac{P[B \mid A]P[A]}{P[B]}$
- In ML, we are interested in $P[\Theta \mid D]$, which reflects our confidence that hypothesis holds given data D, so we have

$$P[\Theta \mid D] = \frac{P[D \mid \Theta]P[\Theta]}{P[D]}$$

- Notice that:

$$P[\Theta \mid D] \propto P[D \mid \Theta]$$

$$P[\Theta \mid D] \propto P[\Theta]$$

$$P[\Theta \mid D] \propto \frac{1}{P[D]}$$

MAP AND MLE

- both MLE and MAP are methods for estimating some variable in the setting of probability distribution or graphical models.
- compute a single estimate rather than a full distribution
- Process: first derive the log likelihood then maximizing it with regard of Θ
- Q: Why working in log space?
A: Logarithm is monotonically increasing, so

$$\arg\max_{\Theta} \log h(\Theta) = \arg\max_{\Theta} h(\Theta)$$

MAXIMUM A POSTERIORI(MAP)

- In many cases, we are interested in finding the most probable hypothesis $h \in H$ given the observed data D.
- MAP hypothesis: maximally probable hypothesis given D
- determine MAP hypothesis using Bayes rule:

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} P[h \mid D] \\
&= \arg\max_{h \in H} \frac{P[D \mid h] P[h]}{P[D]} \\
&= \arg\max_{h \in H} P[D \mid h] P[h] \\
&= \arg\max_{h \in H} \log(P[D \mid h] P[h]) \\
&= \arg\max_{h \in H} \log\left(\prod P[x_i \overset{N}{\underset{i=1}{\mid}} h] \cdot P[h] \right) \\
&= \arg\max_{h \in H} \sum_{i=1}^{N} \log\left(P[x_i \mid h] P[h] \right)
\end{aligned}
$$

* here we drop P[D] because it is a constant independent of h

MAXIMUM LIKELIHOOD ESTIMATION(MLE)

- In some cases, we simply assume $P[h_i] = P[h_j]$ for all $h_i, h_j \in H$
- further simplify the equation and need only consider $P[D \mid h]$
- any hypothesis that maximizes $P[D \mid h]$ is called a maximum likelihood hypothesis

$$
\begin{aligned}
h_{MLE} &= \arg\max_{h \in H} P[D \mid h] \\
&= \arg\max_{h \in H} \log(P[D \mid h]) \\
&= \arg\max_{h \in H} \log(\prod_{i=1}^{N} P[x_i \mid h]) \\
&= \arg\max_{h \in H} \sum_{i=1}^{N} \log(P[x_i \mid h])
\end{aligned}
$$

NAIVE BAYES

- Can be trained with MAP: pick the hypothesis that is most probable

- If the Naive Bayes assumption of conditional independence is satisfied, then this NB classification is identical to MAL classification

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} P[h \mid D] \\
&= \arg\max_{h \in H} \frac{P[D \mid h] P[h]}{P[D]} \\
&= \arg\max_{h \in H} P[D \mid h] P[h] \\
&= \arg\max_{h \in H} \log(P[D \mid h] P[h]) \\
&= \arg\max_{h \in H} \log\left(\prod P[x_i \overset{N}{\underset{i=1}{\mid}} h] \cdot P[h] \right) \\
&= \arg\max_{h \in H} \sum_{i=1}^{N} \log\left(P[x_i \mid h] P[h] \right)
\end{aligned}
$$

MLE PRACTICE

Assume we have a random sample that is Bernoulli distributed $X_1, \ldots, X_n \sim Bernoulli(\Theta)$. We are going to derive the MLE for Θ. Recall that a Bernoulli random variable X takes values in $\{0, 1\}$ and has probability mass function given by

$$P(X; \Theta) = \Theta^X (1 - \Theta)^{1-X}$$

(1) Derive the likelihood, $L(\Theta; X_1, \ldots, X_n)$.

(2) Derive the following formula for the log likelihood:

$$l(\Theta; X_1, \ldots, X_n) = \log(\Theta) \sum_{i=1}^{N} X_i + \log(1 - \Theta)(n - \sum_{i=1}^{N} X_i)$$

(3) Derive the following formula for the MLE:

$$\hat{\Theta} = \frac{1}{n} \sum_{i=1}^{N} X_i$$

MLE VS MAP

1.

[T/F] The value of the Maximum Likelihood Estimate(MLE) is equal to the value of the Maximum A Posteriori(MAP) Estimate with a uniform prior.

True. We know that $P(\Theta|D) \propto P(D|\Theta)P(\Theta)$. The uniform prior gives a constant value on $P(\Theta)$, after proper normalization, we know that likelihood of MLE and the posterior of MAP are the same.

2.

[T/F] The bias of the Maximum Likelihood Estimate(MLE) is typically less than or equal to the bias of the Maximum A Posteriori(MAP) estimate.

True.The MAP estimate injects some prior knowledge and typically adds bias.

NAIVE BAYES PRACTICE

Consider the following data. It has 4 features $\mathbf{x} = (x_1, x_2, x_3, x_4)$ and 3 labels $(+1, 0, -1)$. Assume that the probabilities $p(x_i|y)$ is a Bernoulli distribution and $p(y)$ is a Categorical distribution. Answer the questions that follow under the Naïve Bayes assumption.

x_1	x_2	x_3	x_4	y
1	1	0	1	+1
0	1	1	0	+1
1	0	1	1	0
0	1	1	1	0
0	1	0	0	-1
1	0	0	1	-1
0	0	1	1	-1

Task:

1. Compute the MLE for $P(x_i = 1|y), \forall i \in [1, 4], \forall y \in \{+1, 0, -1\}$.
2. Compute the MLE for the prior probabilities $P(y =_1), P(y = 0), P(y = -1)$
3. Use the values computed in the above two parts to classify the data point $(1, 1, 1, 1)$