BITS F464: Machine Learning

06 Bayesian Learning (MAP and ML)

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Jan 29, 2021 ONLINE (Campus @ BITS-Pilani Jan-May 2021)

http://ktiwari.in/ml

Hypothesis

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	0	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

- In this example h_1, h_2, ... are hypothesis.
- **Hypothesis** is a function that aims to provide value of the Y
- Can you identify h_1 and h_2
- Represent H as candidate set of hypothesis, *i.e.* $h_i \in H$
- To perfectly learn the data, size of H is at least 2^m

Bayesian Learning

It is based on assumption that quantities of interest are governed by probability distribution

- Notation
  - $P(h)$: initial probability that the hypothesis h holds
  - $P(D)$: probability that data D will be observed
  - $P(D|h)$: probability of observing data D given some world in which hypothesis h holds
  - $P(h|D)$: probability of holding hypothesis h when data D is observed

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Maximum a posteriori (MAP)

- Choose a hypothesis that maximizes $P(h|D)$

$$\begin{aligned} h_{MAP} &= \underset{h \in H}{\operatorname{argmax}} P(h|D) \\ &= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)} \\ &= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h) \end{aligned} \quad (1)$$

- Because $P(D)$ is independent of h
- If all the hypothesis are equally probable, we may further simplify called **maximum likelihood (ML)**

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} P(D|h) \quad (2)$$

An Example

Let an illness affects 0.8% of population. There is a test which is 98% accurate for positive and 97% for negative. Consider following two hypothesis

- h_1: person is suffering some illness
- h_2: person is not suffering illness

A **randomly picked** person is tested for illness and is **found positive**. Which is MAP hypothesis out of h_1 and h_2.

- $P(D|h_1)P(h_1) = 0.98 \times 0.008 = 0.0078$ (normalized 0.21)
- $P(D|h_2)P(h_2) = 0.03 \times 0.992 = 0.0298$ (normalized 0.79)

Hypothesis h_2, that is the person is not suffering with illness is most probable.

Let $n = 100000 = (99200 + 800) = (\{96224{+}2976\} + \{16{+}784\})$
$P(h1) =\sim 0.21$ $P(h2) =\sim 0.79$

For our current example

X	Y
10	0
11	0
12	0
13	1
14	0
15	1
16	0
17	1
18	1

h_1	h_2	...
0	1	...
0	0	...
0	1	...
1	1	...
1	1	...
1	0	...
1	1	...
1	0	...
1	1	...

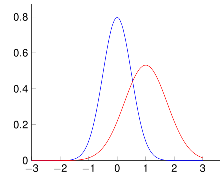
- Let bias for h_1 and h_2 be 2/50 and 6/50
- Since h_1 and h_2 are correct with probability 7/9 and 3/9 respectively
- Posterior is (7/9)*(2/50) and (3/9)*(6/50)
- Normalized probabilities are 0.4375 and 0.5625 respectively
- So MAP hypothesis corresponds to h_2
- Can you guess ML hypothesis? it is h_1

- **Brute-force MAP learning algorithm:** Evaluates posterior probability for all and returns the one with maximum
- **Consistent Learner:** learning algorithm is consistent learner if it provides a hypothesis that commits zero error

Normal Distribution

Many natural phenomenas are assumed to have Normal Distribution

1. Marks obtained by students in a test
2. Weight of a person in the population
3. Sum on dice, tossed 10 times
4. Number of heads in 1000 toss



$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

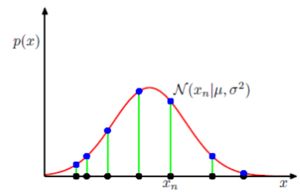
[1]

[1] variance $= \sigma^2$, Standard Deviation $= \sigma$, $\mu + \sigma = 68\%$, $\mu + 2\sigma = 95\%$, $\mu + 2\sigma = 99.7\%$

Maximum Likelihood Estimator of Normal Distribution

- Normal Distribution:

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- $E[x] = \int_{-\infty}^{\infty} N(x|\mu,\sigma^2) \times x\, dx = \mu$
- $E[x^2] = \int_{-\infty}^{\infty} N(x|\mu,\sigma^2) \times x^2\, dx = \mu^2 + \sigma^2$
- $var[x] = E[x^2] - E[x]^2 = \sigma^2$
- Let data set is i.i.d.[2] drawn from normal distribution where true estimates for mean and variance are μ and σ^2
- Likelihood of black points is given by red curve

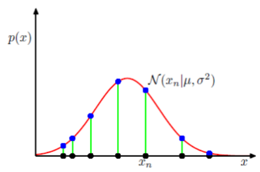
[2] Independent and identically distributed

ML Estimate of Normal Distribution

Assume m data points that are i.i.d. (Independent and identically distributed) in given training data set D

- Probability of the data set D is

$$p(D|\mu,\sigma^2) = \Pi_{i=1}^{m} N(x_i \in D|\mu,\sigma^2) = \Pi_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$



- Changing μ and σ gives different hypothesis h
- $h_{ML} = \text{argmax}_{h \in H} P(D|h)$

Our interest is to determine value of μ and σ that maximizes $p(D|\mu,\sigma^2)$

Maximum Likelihood Estimator of Normal Distribution

MLE would maximize the probability $p(D|\mu,\sigma^2)$ using appropriate μ and σ^2

- What if we optimize log of this probability? (it would be same)

$$\log p(D|\mu,\sigma^2) = \sum_{i=1}^{m} \log N(x_i \in D|\mu,\sigma^2) = \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$
$$= \frac{-1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{1}{2}m.\log(2\pi\sigma^2)$$

- What parameters would optimize this function? differentiate and set to zero (with respect to μ and σ^2). For μ it is

$$\frac{d}{d\mu}\log p(D|\mu,\sigma^2) = \frac{-1}{2\sigma^2}\sum_{i=1}^{m} 2(x_i-\mu)(-1) - 0$$

contd ...

$$\frac{-1}{2\sigma^2}\sum_{i=1}^{m} 2(x_i-\mu)(-1) = 0$$
$$\sum_{i=1}^{m}(x_i-\mu) = 0$$
$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$$

- Similarly for σ^2

$$\frac{d}{d\sigma^2}\log p(\mathbf{x}|\mu,\sigma^2) = \frac{1}{2\sigma^4}\sum_{i=1}^{m}(x_i-\mu)^2 + 0 - \frac{m}{2\sigma^2}$$

contd ...

Equating to zero

$$\frac{1}{2\sigma^4}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{m}{2\sigma^2} = 0$$

Gives

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu)^2$$

For a dataset $\{x_1, x_2, ..., x_m\}$ drawn from normal distribution; one can find maximum likelihood estimates μ and σ^2 as

$$\mu_{ML} = \frac{1}{m}\sum_{i=1}^{m} x_i$$
$$\sigma^2_{ML} = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_{ML})^2$$

Note that

Expected value of μ_{ML} is true value

$$
\begin{aligned}
E[\mu_{ML}] &= E[\frac{1}{m}\sum_{i=1}^{m} x_i] = \frac{1}{m}\sum_{i=1}^{m} E[x_i] \\
&= \frac{1}{m} \times m \times E[x_i] = E[x_i] = \mu
\end{aligned}
$$

Where as expected value of σ^2_{ML} is as follows

$$E[\sigma^2_{ML}] = \frac{m-1}{m}\sigma^2$$

MLE is biased in case of variance but not for mean

Note that

For convenient notation, let us write μ_{ML} as $\bar{x}$

$$
\begin{aligned}
E[\sigma^2_{ML}] &= E[\frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2] = E[\frac{1}{m}\sum_{i=1}^{m}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)] \\
&= \frac{1}{m}E[\sum_{i=1}^{m} x_i^2 - 2\bar{x}\sum_{i=1}^{m} x_i + \sum_{i=1}^{m}\bar{x}^2] = \frac{1}{m}E[\sum_{i=1}^{m} x_i^2 - 2\bar{x}(m.\bar{x}) + m.\bar{x}^2] \\
&= \frac{1}{m}E[\sum_{i=1}^{m} x_i^2 - m.\bar{x}^2] = \frac{1}{m}[m.E[x^2] - m.E[\bar{x}^2]] = E[x^2] - E[\bar{x}^2]
\end{aligned}
$$

Since $\sigma_x^2 = E[x^2] - E[x]^2$ $\quad \sigma_{\bar{x}}^2 = E[\bar{x}^2] - E[\bar{x}]^2$ $\quad E[x] = E[\bar{x}] = \mu$

$$E[\sigma^2_{ML}] = (\sigma_x^2 + E[x]^2) - (\sigma_{\bar{x}}^2 + E[\bar{x}]^2) = \sigma_x^2 - \sigma_{\bar{x}}^2$$

Since $\sigma_{\bar{x}}^2 = var[\bar{x}] = var[\frac{1}{m}\sum_{i=1}^{m} x_i] = \frac{1}{m^2}var[\sum_{i=1}^{m} x_i] = \frac{1}{m^2}\sum_{i=1}^{m} var[x_i] = \frac{1}{m^2}m.var[x_i] = \frac{1}{m}var[x_i] = \frac{1}{m}\sigma_x^2$

$$E[\sigma^2_{ML}] = \sigma_x^2 - \frac{1}{m}\sigma_x^2 = \frac{m-1}{m}\sigma_x^2$$

To overcome this, we could multiply sample variance with $m/(m-1)$

Frequentist vs Bayesian Approach

Observed data D, is coming from some unknown distribution $p(x, y)$

- Can we get parameters θ of some known distribution $p_\theta(x, y)$ to match the unknown distribution $p(x, y)$ with high probability

MLE: Maximum Likelihood Estimation

$$\theta = \underset{\theta}{\text{argmax}}\, p(D)$$

- Get a θ that maximizes the probability of data
- θ is a parameter
- **Frequentist Approach**

MAP: Maximum Aposteriori Probability

$$\theta = \underset{\theta}{\text{argmax}}\, p(\theta|D)$$

- Get a θ that has maximum probability given the data
- θ is a random variable
- **Bayesian Approach**

Bayesian approach have to define some prior distribution over the θ

Bayesian classification is given by $p(y|x) = \int_\theta p(y|\theta) \times p(\theta|d)\, d\theta$

Thank You!

Thank you very much for your attention!

Queries ?