innovate achieve lead

BITS Pilani
Pilani Campus

M.Tech DSE
Machine Learning
(DSECL ZG565)

Dr. Monali Mavani

Part – I Agenda

Bayesian learning

- Bayes Theorem (T1 book by Tom Mitchell - 6.2)
- MAP Hypothesis (T1 book by Tom Mitchell - 6.3)
- MLE Hypothesis (T1 book by Tom Mitchell - 6.4)

Probability Distributions

- The outcomes for random variables and their associated probabilities can be organized in to distributions
- Two types of distributions based on types of Random variables: Discrete and Continuous
- Discrete:
  - Binomial, Poisson, Geometric distributions
- Continuous
  - Gaussian, exponential, t, F, chi-squared distributions

Describing distributions

- One way is to construct a graph and analyze the graph to make inferences
  - Discrete: Prob Mass Function (pmf), Cumulative density function
  - Continuous: prob density function (pdf), Cumulative density function
- Mean, variance and standard deviations to represent the entire distribution

JOINT Distributions

- Probability distribution of two random variables X $\{x_1, x_2, \ldots x_n\}$ and Y$\{y_1, y_2..y_k\}$
  - Occurrence of X=xi and Y=yi together
- Example:
  - P(X=0, Y<=1)
  - P(X=1)

    $= \sum_{y=0}^{2} P(X=1, Y)$

    = 1/6 + 1/6 + 1/8

		Y = 0	Y = 1	Y = 2
X	0	1/4	1/6	1/8
	1	1/6	1/6	1/8

Estimate Probabilities from Data

- For continuous attributes:
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, use it to estimate the conditional probability $P(X_i|Y)$

Estimate Probabilities from Data

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Normal distribution:

$$P(X_i \mid Y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(X_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

  – One for each (X_i, Y_i) pair

- For (Income, Class=No):
  – If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Parameters and Parametric Models

Distribution	Parameters
Bernoulli(p)	$\theta = p$
Poisson(λ)	$\theta = \lambda$
Uniform(a,b)	$\theta = (a, b)$
Normal(μ, σ^2)	$\theta = (\mu, \sigma^2)$
Y = mX + b	$\theta = (m, b)$

Usually refer to parameters of distribution as θ

Note that θ that can be a vector of parameters

Likelihood

innovate achieve lead

- Consider IID random samples $X_1, X_2, \ldots, X_n$ where X_i is a sample from the density function $f(X_i | \theta)$.
- we define the likelihood of our data given parameters θ :

$$L(\theta) = \prod_{i=1}^{n} f(X_i | \theta)$$

- Intuitively: what is probability of observed data using density function $f(Xi \mid \theta)$, for some choice of θ. The density of X depends on its parameters, θ

If X is discrete,

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i \mid \theta)$$

If X is continuous,

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

Maximum Likelihood Estimation (MLE)

- MLE: to chose values of our parameters (θ) that maximizes the likelihood function i.e the best choice of values for our parameters. Formally,

$$\boxed{\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)}$$

- Log Likelihood

$$\boxed{LL(\theta) = \log L(\theta)}$$

- **If the sample is large, MLE will yield an excellent estimator of θ.**

- **MLE answers the question: For which parameter value does the observed data have the biggest probability?**

Bernoulli MLE Estimation

Consider IID random variables $X_1, X_2, \ldots, X_n$ where $X_i \sim Ber(p)$.

PMF of a Bernoulli $\boxed{p^{X_i}(1-p)^{1-X_i}}$

Remember: Some terminology

- Likelihood function: $P(\text{data} \mid \theta)$

- Prior: $P(\theta)$

- Posterior: $P(\theta \mid \text{data})$

Bayes Theorem

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

- $P(h)$ = prior probability of hypothesis h
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

MAP Hypothesis

- **Machine learning** is interested in the best hypothesis *h* from some space H, given observed training data D
- best hypothesis ≈ most probable hypothesis
- Bayes Theorem provides a direct method of calculating the probability of such a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself

MAP Hypothesis

- in many learning scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis h ∈ H given the observing training data D
- any maximally probable hypothesis is called maximum a posteriori (MAP) hypotheses

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

- Note that P(D) can be dropped, because it is constant independent of h

BITS Pilani, Pilani Campus

ML Hypothesis

- When no prior information is available, all hypothesis are equally likely i.e. p(hi) = p(hj)
  - This is also true for a balanced class problem where all the classes are equally likely
  - This is known as Uniform prior
  - MAP hypothesis further simplifies to:

$$H_{ML} = argmax_{h \in H} \mathrm{P(D|h)}$$

This is called Maximum Likelihood Hypothesis

$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i)$$

Note that in this case P(h) can be dropped, because it is equal for every h ∈ H

Brute Force MAP Hypothesis

1. For each hypothesis *h* in *H*, calculate the posterior probability

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

2. Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}}\, P(h|D)$$

ML setting

- Bayesian Analysis
  - start with some belief about the system, called a prior.
  - Then we obtain some data and use it to update our belief.
  - The outcome is called a posterior.
  - Should we obtain even more data, the old posterior becomes a new prior and the cycle repeats.
  - People often use likelihood for evaluation of models: a model that gives higher likelihood to real data is better

ML Setting

- P(h | D) a posterior determines the class label
- It's a probability distribution over model parameters obtained from prior beliefs and data.
- When one uses likelihood to get point estimates of model parameters, it's called Maximum Likelihood estimation or MLE.
- If one also takes the prior into account, then it's maximum a posteriori estimation (MAP).
- MLE and MAP are the same if the prior is uniform
- This forms the basis for Naïve Bayes classifier

Example MLE

Example 1: Suppose that X is a discrete random variable with the following probability mass function: where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observations

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

were taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of θ.

Example MAP

1. Example on MAP algorithm:

Let X be continuous random variable with probability density function P(X) given by:

$$f(x) = \begin{cases} 2x, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

Given another distribution $p(Y|X = x) = x(1-x)^{y-1}$ Find MAP estimate of X given Y=3

Least-Squared Error

- If *y* is continuous:
  - Sum-of-Squared-Differences (SSD) error between predicted and true *y*:

$$\mathbf{E} = \sum_{\mathbf{i}=1}^{\mathbf{n}} (\mathbf{f}(\mathbf{x}_{\mathbf{i}}) - \mathbf{y}_{\mathbf{i}})^2$$

Bayesian justification to Least-Squared Error

- Problem: learning continuous-valued target functions
- Minimizing the sum of squared errors
- E.g linear regression, NN, Polynomial curve fitting
- **under certain assumptions any learning algorithm that minimizes the squared error between the output hypothesis and the training data, will output a ML hypothesis**

BITS Pilani, Pilani Campus

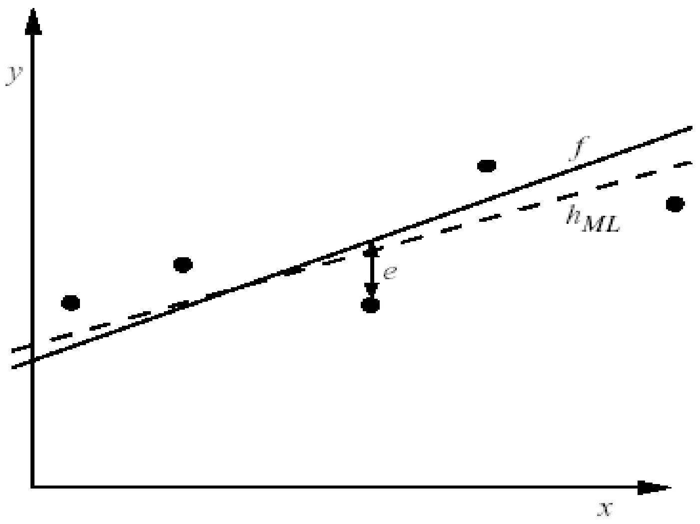
Learning A Real Valued Function

- Problem setting:
  - ✓ $(\forall h \in H)\ [\ h : X \rightarrow \Re\ ]$ and training examples of the form $<x_i\ ,\ d_i>$
  - ✓ unknown target function $f : X \rightarrow \Re$
  - ✓ Training examples $<x_i, d_i>$, where d_i is noisy training value
  - ✓ $d_i = f(x_i) + e_i$
  - ✓ e_i is random variable (noise) drawn independently for each x_i according to some Gaussian distribution with mean=0

Learning A Real Valued Function: CASE of Linear Regression loss

Then the maximum likelihood hypothesis h_{ML} is the one that minimizes the sum of squared errors:

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$



Learning A Real Valued Function: CASE of Linear Regression loss

$$h_{ML} = \underset{h \in H}{argmax}\; p(D|h)$$

- The training examples are assumed to be mutually independent given *h*

$$h_{ML} = \underset{h \in H}{argmax} \prod_{i=1}^{m} p(d_i|h)$$

- Given the noise e_i obeys a Normal distribution with zero mean and unknown variance σ , each d_i must also obey a Normal distribution around the true target value $f(x_i)$
- Because we are writing the expression for d_i given that h is correct description of target function *f*. We will also substitute, $\mu = f(x_i) = h(x_i)$. Hence:

$$h_{ML} = \underset{h \in H}{argmax} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i - h(x_i))^2}$$

BITS Pilani, Pilani Campus

Maximum Likelihood and Least-Squared Error

- It is common to maximize the less complicated logarithm, which is justified because of the monotonicity of this function

$$h_{ML} = \underset{h \in H}{argmax} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

- The first term in this expression is a constant independent of *h* and can therefore be discarded

$$h_{ML} = \underset{h \in H}{argmax} \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

- Maximizing this negative term is equivalent to minimizing the corresponding positive term

$$h_{ML} = \underset{h \in H}{argmin} \sum_{i=1}^{m} \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

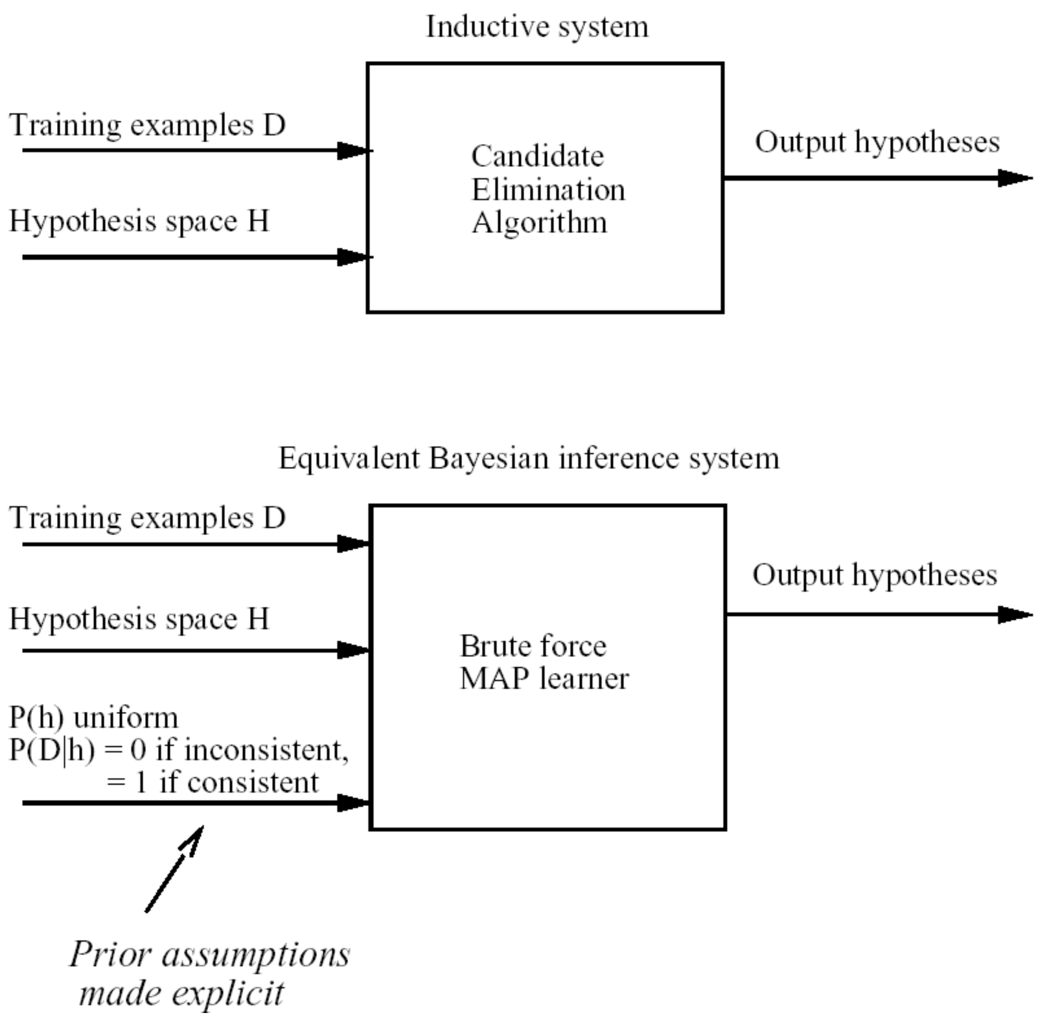
Maximum Likelihood and Least-Squared Error

- Finally, all constants independent of *h* can be discarded

$$h_{ML} = \underset{h \in H}{argmin} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

- h_{ML} is one that minimizes the sum of the squared error

Characterizing Learning Algorithms by Equivalent MAP Learners



Some Additional References

https://web.stanford.edu/class/archive/cs/cs109/cs109.1166//handouts/overview.html

https://www.cs.cmu.edu/~ninamf/courses/601sp15/lectures.shtml

Practice Problem

Example – use Bayes Rule

- Consider a medical diagnosis problem in which there are two alternative hypothesis
  - ✓The patient has particular form of cancer
  - ✓The patient does not
- The available data is from particular laboratory with two possible outcomes: $\oplus$ (positive) and $\ominus$ (negative)

$$P(cancer) = .008 \qquad P(\neg cancer) = 0.992$$
$$P(\oplus|cancer) = .98 \qquad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03 \qquad P(\ominus|\neg cancer) = .97$$

- Suppose a new patient is observed for whom the lab returns a positive ($\oplus$) result
- Should we diagnosis the patient as having a cancer or not?