# CS5010

# Lecture 5, Statistical Learning, Chapter 20 sections 1 - 3

November 29, 2021

---

## Contents

# 1 Outline

- Bayesian learning
- Maximum a posteriori and maximum likelihood learning
- Bayes net learning
  - ML parameter learning with complete data
  - Linear regression

Most decisions to be made in AI are probabilistic in nature. Need framework for reasoning and this is where Bayes theorem comes in.

Calculate the probability of each hypothesis given the data. Use ML to make predictions on that basis.

**EH**: this lecture is covering Bayesian learning and provides some basic examples. In my opinion Lei's notes and section of the course cover these areas in much greater detail and are a lot clearer.

# 2 Full Bayesian Learning

## 2.1 Description

View learning as Bayesian updating of a probability distribution over the hypothesis space. In other words this is different ways of explaining something that we've observed.

$H$ is the hypothesis variable, values $h_1, h_2, ...$ prior $P(H)$

$j$th observation $d_j$ gives the outcome of random variable $D_j$ training data $\mathbf{d} = d_1, ..., d_n$

Given the data so far, each hypothesis has a posterior probability:

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

where $P(\mathbf{d}|h_i)$ is called the likelihood

Posterior probability of the hypothesis given the evidence = prior prob. of the hypothesis times the likelihood function.

The alpha term is the normalising constant that accounts for the probability that the hypothesis does not apply to the evidence.

Predictions use a likelihood-weighted average over the hypotheses:

$$P(X|\mathbf{d}) = \sum_i P(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \sum_i P(X|h_i)P(h_i|\mathbf{d})$$

No need to pick one best-guess hypothesis!

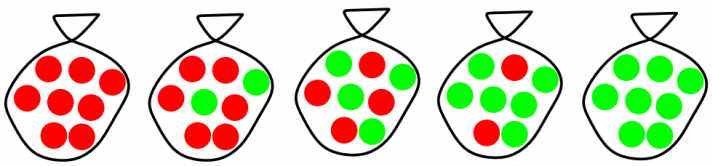
Notes

- As an example, Oggie references AI masters student approaching about doing a project. He assumes an appropriate prior probability that the person will do a good job is 60%. He then looks up their background and observes that they have a fine arts degree, this would initially update probabilities downwards, posterior is then 10%. He then spoke with the person and based on the discussion sees they are intelligent and a fast learner, updated probability is then 90%.
- Probability evolves as evidence is gathered. Bayesian statistics is a means of updating the probability.
- Emboldened **d** in these formulae indicates that it is a vector.

## 2.2 Example - Candies in a bag

Suppose there are five kinds of bags of candies:

- 10% are $h_1$: 100% cherry candies
- 20% are $h_2$: 75% cherry candies + 25% lime candies
- 40% are $h_3$: 50% cherry candies + 50% lime candies
- 20% are $h_4$: 75% cherry candies + 25% lime candies
- 10% are $h_5$: 100% lime candies



Taking $X$ to be the probability that the next candy is a lime, we then have:

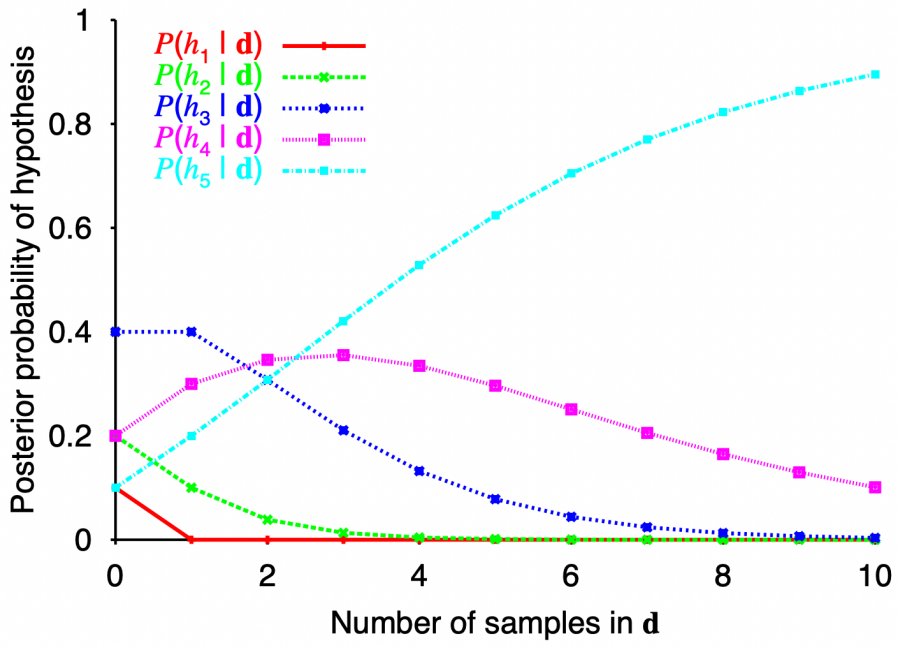
- $P(X|h_1) = 0$
- $P(X|h_2) = 0.25$
- $P(X|h_3) = 0.5$
- $P(X|h_4) = 0.75$
- $P(X|h_5) = 1$

Suppose we draw 10 green candies in a row from the bag. What kind of bag is it? What flavour will the next candy be?

Notes

Problem with this example is that it doesn't state the number of candies in each bag and whether they are replaced or not. Drawing candies would impact the proportion in each bag. We assume that there are an infinite number of candies in each bag.

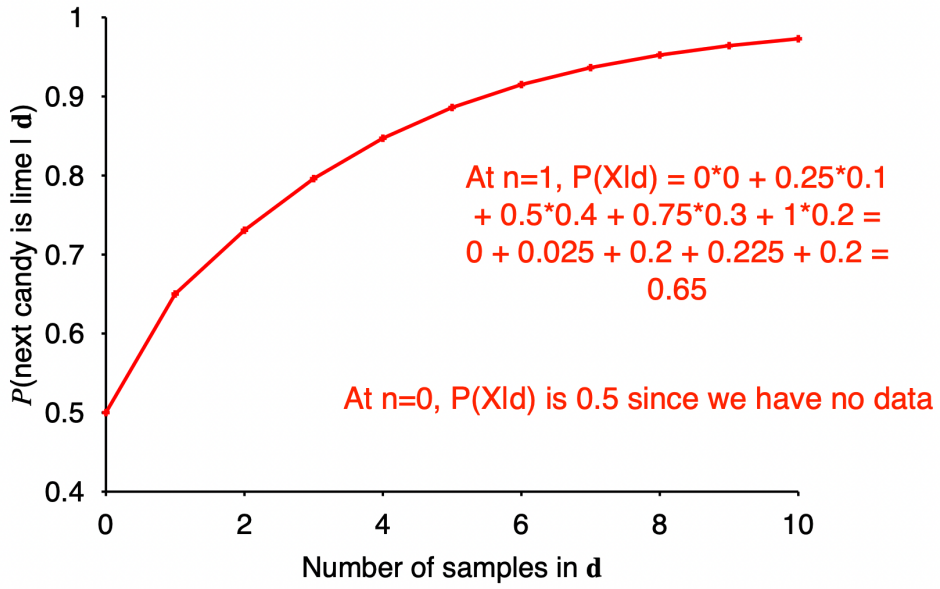
### 2.3 Posterior probability of hypotheses




- This graph shows the probability that we are drawing from each bag as we draw more and more green candies.

- Consequently the probability that we are drawing from $h_5$ grows as $d$ gets larger.
- Red line drops immediately to zero as we know this bag contains no lime candies.
- The green, dark blue and pink lines will asymptotically approach 0 as $d$ gets larger.
- The lines start at different points as the number of bags isn't uniformly distributed and there are different proportions of each bag. This is outlined in the problem statement.

### 2.4 Prediction probability




The probability that the correct hypothesis is chosen tends towards one as the sample grows larger.

## 3 MAP Approximation

Summing over the hypothesis space is often intractable (for example, in this case there are $2^{2^6}$ = 18,446,744,073,709,551,616 Boolean functions of 6 attributes)

Maximum a posteriori (MAP) learning: choose $h_{MAP}$ maximising $P(h_i|\mathbf{d})$

I.e. maximise $P(\mathbf{d}|h_i)P(h_i)$ or $logP(\mathbf{d}|h_i) + logP(h_i)$

Log terms can be viewed as (negative of) bits to encode data given hypothesis + bits to encode hypothesis This is the basic idea of minimum description length (MDL) learning. MDL is not

examinable.

For deterministic hypotheses, $P(\mathbf{d}|h_i$ is 1 if consistent, 0 otherwise $\implies$ MAP = simplest consistent hypothesis (cf. science)

MAP is dangerous with sparse evidence. In the example, we'd select $h_5$ after three lime candies with probability 1. The actual Bayesian prediction has P = 0.8

As more data arrives the MAP & Bayesian probabilities converge

Notes

- This is effectively a formalisation of Occam's razor, given two hypotheses that explain the same data, we should favour the less complex one.
- Update prior probability and choose hypothesis which best explains the data.
- MDL not examinable.

## 4 ML Approximation

For large data sets, prior becomes irrelevant.

Maximum likelihood (ML) learning: choose $h_{ML}$ maximising $P(\mathbf{d}|h_i)$

I.e. simply get the best fit to the data; identical to MAP for uniform prior (which is reasonable if all hypotheses are of the same complexity).

ML is the "standard" (non-Bayesian) statistical learning method

We assume a uniform prior over the hypothesis space. No hypothesis is preferred before we start seeing data. Enough data will swamp this prior distribution.

Notes

- Trying to determine which hypothesis best determines the data.
- Standard frequentist interpretation of data.

## 5 ML Parameter learning in Bayes nets

**Lecture notes here do not really discuss Bayes nets, better to refer to Lei's notes for a proper breakdown.**

Bags from a new manufacturer; fraction $\theta$ of cherry candies?

Any $\theta$ is possible: continuum of hypotheses $h_\theta$

$\theta$ is a parameter for this simple (binomial) family of models.

Suppose we unwrap N candies, c cherries and $l = N - c$ limes

These are i.i.d. (independent, identically distributed) observations, so

$$P(\mathbf{d}|h_\theta) = \prod_{j=1}^{N} P(d_j|h_\theta) = \theta^c(1-\theta)^l$$

Maximise this w.r.t. $\theta$ - which is easier for the log-likelihood:

$$L(\mathbf{d}|h_\theta) = logP(\mathbf{d}|h_\theta) = \sum_{j=1}^{N} logP(d_j|h_\theta) = c * log\theta + l * log(1-\theta)$$

$$\frac{dL(\mathbf{d}|h_\theta)}{d\theta} = \frac{c}{\theta} - \frac{l}{1-\theta} = 0 \implies \theta = \frac{c}{c+l} = \frac{c}{N}$$

Seems sensible, but causes problems with 0 counts!

Maxima and minima occur at places where the derivative is zero

Method:

1. Derive expression for likelihood of data as a function of the parameters.
2. Obtain derivative of the log likelihood wrt each parameter.
3. Find parameter values where derivatives are zero.

Notes

- This is an example of Bayesian inference, where we are trying infer the value of some hidden or unknown variable. In this case the proportion of cherries in the bag.
- Key consideration is how we model a process like this.
- Result above for $\theta$ is quite intuitive if we take a step back.
- Example of binomial distribution.

## 5.1 Multiple parameters

Red/green wrapper depends probabilistically on flavor:

Likelihood for, e.g., cherry candy in green wrapper:

$$\begin{aligned} &P(F = cherry, W = green|h_{\theta,\theta_1,\theta_2}) \\ &= P(F = cherry|h_{\theta,\theta_1,\theta_2})P(W = green|F = cherry, h_{\theta,\theta_1,\theta_2}) \\ &= \theta(1-\theta_1) \end{aligned}$$

N candies, $r_c$ red-wrapped cherry candies, etc.:

$$P(\mathbf{d}|h_{\theta,\theta_1,\theta_2}) = \theta^c(1-\theta)^l * \theta_1^{r_c}(1-\theta_1)^{g_c} * \theta_2^{r_l}(1-\theta_2)^{g_l}$$

$$\begin{aligned} L = &[clog\theta + l * log(1-\theta)] \\ &+[r_c log\theta_1 + g_c log(1-\theta_1)] \\ &+[r_l log\theta_2 + g_l log(1-\theta_2)] \end{aligned}$$

Derivatives of L contain only the relevant parameter:

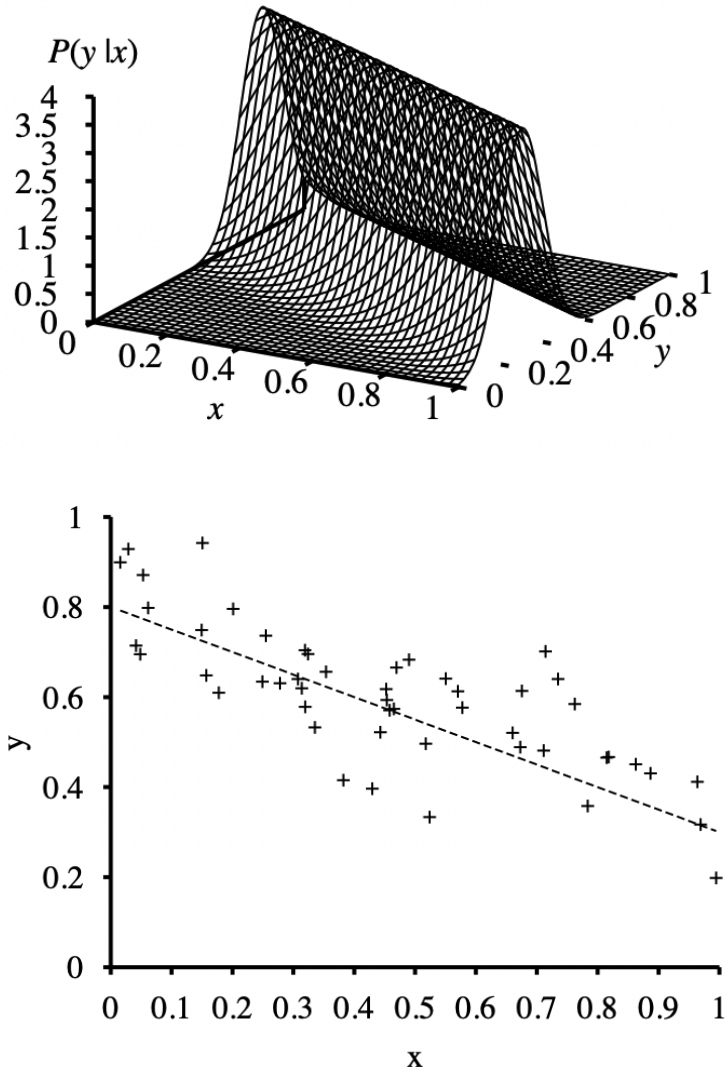
$$\begin{aligned} \frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{l}{1-\theta} = 0 \qquad &\implies \theta = \frac{c}{c+l} \\ \frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \qquad &\implies \theta_1 = \frac{r_c}{r_c+g_c} \\ \frac{\partial L}{\partial \theta_2} = \frac{r_l}{\theta_2} - \frac{g_l}{1-\theta_2} = 0 \qquad &\implies \theta_2 = \frac{r_l}{r_l+g_l} \end{aligned}$$

With complete data, parameters can be learned separately

Notes

- This is an extension of drawing from the bag and is meant as an extension to that problem.
- Again we define the likelihood function, convert to log likelihood and then take partial derivatives to determine the maximum estimators.
- Taking a step back, again the results are somewhat intuitive.

## 5.2 Example: linear Gaussian model




Maximising $P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(\theta_1 x+\theta_2))^2}{2\sigma^2}}$ w.r.t. $\theta_1$, $\theta_2$

= minimising $E = \sum_{j=1}^{N} (y_j - (\theta_1 x_j + \theta_2))^2$

That is, minimising the sum of squared errors (difference between each data point and your regression line) gives the ML solution for a linear fit assuming Gaussian noise of fixed variance

Continuous models use the same methods and principles as for the discrete case. The maths is trickier though

Notes

- Again, define likelihood function and optimise based on that.

## 6 Summary

Full Bayesian learning gives best possible predictions but is intractable.

MAP learning balances complexity with accuracy on training data

Maximum likelihood assumes uniform prior, OK for large datasets

1. Choose a parameterized family of models to describe the data requires substantial insight and sometimes new models
2. Write down the likelihood of the data as a function of the parameters may require summing over hidden variables, i.e. inference
3. Write down the derivative of the log likelihood w.r.t. each parameter
4. Find the parameter values such that the derivatives are zero may be hard/impossible; modern optimisation techniques help