CS 760: Machine Learning Fall 2020

Topic 10: Bayesian Learning

Instructor: Daniel L. Pimentel-Alarcón

GO GREEN. AVOID PRINTING, OR PRINT 2-SIDED OR MULTIPAGE.

10.1 Review of Bayes Rule

Definition 10.1 (Conditional probability)**.** Let A, B be two events. The *conditional probability* that A occurs given B occurred is

$$\mathbb{P}(A|B) \ := \ \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Example 10.1. Consider a 6-faced fair die. Let $A = \{1, 2\}$ be the event that the die rolls either 1 or 2, and similarly for $B = \{2, 3\}$. The probability that A occurs is $\mathbb{P}(A) = 1/3$. However, if you already know that B occurred, then the conditional probability that A also ocurred increases to

$$\mathbb{P}(A|B) \ = \ \frac{1/6}{1/3} \ = \ \frac{1}{2}.$$

Given the conditional probability $\mathbb{P}(A|B)$, Bayes rule gives us a formula for the *posterior* probability, $\mathbb{P}(B|A)$.

Lemma 10.1 (Bayes rule)**.** Let A, B be two events. Then

$$\mathbb{P}(B|A) \ = \ \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \tag{10.1}$$

Bayes rule plays a crucial role in modern applications, as it establishes a clear relationship between conditional and posterior probabilities.

Example 10.2. Geneticists have determined that 90% of the people with disease B have gene A active, i.e., $\mathbb{P}(A|B) = 0.9$. If you sequence your genome and find out that your gene A is active, what is the probability that you develop disease B? In other words, what is $\mathbb{P}(B|A)$? At first glance you might think it is very likely that you will develop disease B. However, to determine this you need to know $\mathbb{P}(A)$ and $\mathbb{P}(B)$. Of the whole population, if only 5% have disease B, while 45% have gene A active,

what is $\mathbb{P}(B|A)$? This is a simple application of Bayes rule:

$$\mathbb{P}(B|A) \;=\; \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \;=\; \frac{(0.9)(0.05)}{0.45} \;=\; 0.1$$

Definition 10.2 (Independent events)**.** Let A, B be two events. We say A and B are *independent* if

$$\mathbb{P}(A|B) = \mathbb{P}(A).$$

Example 10.3. Consider two fair dice. Let A be the event that the first die is 1; let B be the event that the second die is 1. Then

$$\mathbb{P}(A|B) \;=\; \frac{\mathbb{P}(A=1 \cap B=1)}{\mathbb{P}(B=1)} = \frac{1/36}{1/6} = \frac{1}{6} = \mathbb{P}(A).$$

Hence the events A and B are independent. This matches our intuition that one die has no influence on the outcome of the other.

10.2 Naive Bayes

Naive Bayes is one of the simplest supervised classification methods: one has a collection of N training pairs $\{(\mathbf{x}_\mathrm{i}, \mathrm{y}_\mathrm{i})\}_{\mathrm{i}=1}^{\mathrm{N}}$, where $\mathbf{x}_\mathrm{i} \in \mathbb{R}^\mathrm{D}$ contains the D features of the i^{th} sample (e.g., glucose level, height, gender, etc.), and $\mathrm{y}_\mathrm{i} \in \{1, \ldots, \mathrm{C}\} =: [\mathrm{C}]$ denotes the class to which sample i belongs (e.g., healthy or diabetic). Given a new sample $\mathbf{x}$, the goal is to determine its class y. The main idea behind Naive Bayes is to choose the class with the highest *posterior* probability under the *naive* assumption that features are independent. That is,

$$\begin{aligned}
\hat{\mathrm{y}} \;&:=\; \arg\max_{\mathrm{y}\in[\mathrm{C}]} \; \mathbb{P}(\mathrm{y}|\mathbf{x}) && (10.2)\\
&=\; \arg\max_{\mathrm{y}\in[\mathrm{C}]} \; \frac{\mathbb{P}(\mathbf{x}|\mathrm{y})\mathbb{P}(\mathrm{y})}{\mathbb{P}(\mathbf{x})} && (10.3)\\
&=\; \arg\max_{\mathrm{y}\in[\mathrm{C}]} \; \mathbb{P}(\mathbf{x}|\mathrm{y})\mathbb{P}(\mathrm{y}) && (10.4)\\
&=\; \arg\max_{\mathrm{y}\in[\mathrm{C}]} \; \mathbb{P}(\mathrm{y}) \prod_{\mathrm{j}=1}^{\mathrm{D}} \mathbb{P}(\mathrm{x}_\mathrm{j}|\mathrm{y}). && (10.5)
\end{aligned}$$

Here $\mathbb{P}(\mathrm{y}|\mathbf{x})$ in (10.2) is the posterior probability that we aim to maximize. (10.3) decomposes the posterior into the *prior* probability of a class $\mathbb{P}(\mathrm{y})$, and the *conditional* probability $\mathbb{P}(\mathbf{x}|\mathrm{y})$ of the sample given each class (a.k.a. likelihood), using Bayes rule. (10.4) follows because $\mathbb{P}(\mathbf{x})$ does not depend on y, and (10.5) by the independence assumption of the features of $\boldsymbol{x}$, denoted by x_j. With (10.5) it all boils down to estimating $\mathbb{P}(\mathrm{y})$ and $\mathbb{P}(\mathrm{x_j}|\mathrm{y})$ for each class y, which can be done by maximum likelihood using the given training data.

Example 10.4 (Bernoulli)**.** Congratulations. You have just been hired by Google to develop a new email spam filter using Naive Bayes. To this end you can use the following dataset, indicating the words included in a list of emails:

			Samples (Emails)									
			1	2	3	4	5	6	7	8	9	10
Features	1	congratulations	1	1	1	1	1	0	0	0	0	1
	2	you	1	1	1	0	0	0	1	1	0	0
	3	won	0	1	1	1	1	1	0	0	0	1
	4	free	1	1	1	1	1	1	1	0	0	0
	5	gift	0	0	1	1	1	1	0	1	0	0
	6	attached	0	0	1	0	0	0	1	1	1	0
	7	sincerely	1	0	1	0	0	1	0	0	1	1
	8	thanks	0	1	0	1	1	0	1	1	0	0
y		**Class**	**Spam**						**Not Spam**			

Given a new email saying "congratulations, you won free gift", the goal is to determine whether it is spam or not. To this end, Naive Bayes first estimates $\mathbb{P}(y)$ using maximum likelihood. Recall that the likelihood of i.i.d. Bernoulli(p) samples $\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{N}$ is

$$\mathbb{P}(\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{N}|\mathrm{p}) \;=\; \mathrm{p}^{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y}_\mathrm{i}} (1-\mathrm{p})^{\mathrm{N}-\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y}_\mathrm{i}},$$

so the *maximum likelihood estimate* (MLE) is

$$\hat{\mathrm{p}} \;:=\; \arg\max_{\mathrm{p}} \;\; \mathrm{p}^{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y}_\mathrm{i}} (1-\mathrm{p})^{\mathrm{N}-\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y}_\mathrm{i}} \;=\; \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y}_\mathrm{i},$$

where the last step can be derived using our Optimization 101 recipe (take log, take derivative, set to zero, and solve). We thus conclude that the MLE of $\mathbb{P}(y)$ is given by the fraction of samples that fall under each class:

$$\begin{aligned} \hat{\mathbb{P}}(y = \text{spam}) \;&=\; \frac{6}{10} \;=\; \frac{3}{5}, \\ \hat{\mathbb{P}}(y = \text{not spam}) \;&=\; \frac{4}{10} \;=\; \frac{2}{5}. \end{aligned}$$

Similarly, for each feature x_j we can obtain the MLE of its probability $\mathbb{P}(x_\mathrm{j} = 1|y)$ as the fraction of 1's in each class,

$$\hat{\mathbb{P}}(x_\mathrm{j} = 1|y = \mathrm{y}) \;=\; \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{1}_{\{\mathrm{x_{ij}}=1, \mathrm{y_i}=\mathrm{y}\}}}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{1}_{\{\mathrm{y_i}=\mathrm{y}\}}}, \tag{10.6}$$

to obtain the following conditional probability matrix estimate:

			$\hat{\mathbb{P}}(x_\mathrm{j} = 1\|y)$	
			Spam	**Not Spam**
Features	1	congratulations	$5/6$	$1/4$
	2	you	$1/2$	$1/2$
	3	won	$5/6$	$1/4$
	4	free	1	$1/4$
	5	gift	$2/3$	$1/4$
	6	attached	$1/6$	$3/4$
	7	sincerely	$1/2$	$1/2$
	8	thanks	$1/2$	$1/2$

Given our new sample with feature vector $\mathbf{x} = [\ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 0\ \ 0\ \ 0\ ]^\mathsf{T}$, all we have to do is estimate the posterior probability of each class under the naive assumption:

$$\begin{aligned}\hat{\mathbb{P}}(y = \text{spam}|\mathbf{x}) \ &\propto\ \hat{\mathbb{P}}(y = \text{spam}) \cdot \prod_{\text{j}=1}^{\text{D}} \hat{\mathbb{P}}(\text{x}_\text{j}|y = \text{spam}) \\ &= \ \frac{3}{5} \cdot \hat{\mathbb{P}}(x_1 = 1|y = \text{spam}) \cdot \hat{\mathbb{P}}(x_2 = 1|y = \text{spam}) \cdot \hat{\mathbb{P}}(x_3 = 1|y = \text{spam}) \\ &\quad\ \hat{\mathbb{P}}(x_4 = 1|y = \text{spam}) \cdot \hat{\mathbb{P}}(x_5 = 1|y = \text{spam}) \cdot \hat{\mathbb{P}}(x_6 = 0|y = \text{spam}) \\ &\quad\ \hat{\mathbb{P}}(x_7 = 0|y = \text{spam}) \cdot \hat{\mathbb{P}}(x_8 = 0|y = \text{spam}) \\ &= \ \frac{3}{5} \cdot \frac{5}{6} \cdot \frac{1}{2} \cdot \frac{5}{6} \cdot 1 \cdot \frac{2}{3} \cdot \frac{5}{6} \cdot \frac{1}{2} \cdot \frac{1}{2} \ = \ 0.0289.\end{aligned}$$

Similarly, we can estimate $\hat{\mathbb{P}}(y = \text{not spam}|\mathbf{x})$, and choose the class with the largest posterior probability. In this case, how would you classify the new sample $\mathbf{x}$? Notice that if a feature x_j never takes a given value x in class y, then $\mathbb{P}(x_\text{j} = \text{x}|y = \text{y})$ will be equal to zero, which in turn would also zero out the posterior $\mathbb{P}(y = \text{y}|\mathbf{x})$, neglecting the information in all other probabilities involved. To avoid this issue, the estimation of $\mathbb{P}(x_\text{j} = \text{x}|y = \text{y})$ is often *smoothed* by adding a *regularization* term (representing a *virtual* sample with feature $x_\text{j} = \text{x}$ in class y), so that no probability is ever estimated to be exactly zero. For example, instead of (10.6) we could use:

$$\hat{\mathbb{P}}(x_\text{j} = 1|y = \text{y}) \ = \ \frac{\sum_{\text{i}=1}^{\text{N}} \mathbb{1}_{\{\text{x}_\text{ij}=1, \text{y}_\text{i}=\text{y}\}} + 1}{\sum_{\text{i}=1}^{\text{N}} \mathbb{1}_{\{\text{y}_\text{i}=\text{y}\}} + \text{K}_\text{j}},$$

where K_j is the number of possible values that the feature x_j may take. The case when we add 1 sample per feature value per class is called *Laplace smoothing*, and the general case is called *Lidstone smoothing*.

Example 10.5 (Gaussian)**.** Suppose you are the detective in charge of a murder, and have been able to gather the following evidence about the killer: shoe size = 42cm, height = 180cm, and maximum running speed = 5.5 minutes/mile. To narrow down your list of suspects you first want to determine whether the killer is male or female, using a Naive Bayes approach. To this end you may use the following information:

			Samples (People)					
			1	2	3	4	5	6
Features	1	Shoe size (cm)	41	43	44	45	37	39
	2	Height (cm)	170	175	185	180	160	170
	3	Max Speed (min/mile)	6	7	6.5	7.5	6.5	7
y		**Gender**	**Male**				**Female**	

First we obtain the MLE of the priors as the fraction of samples in each class. In this case:

$$\begin{aligned}\hat{\mathbb{P}}(y = \text{male}) \ &= \ \frac{4}{6} \ = \ \frac{2}{3}, \\ \hat{\mathbb{P}}(y = \text{female}) \ &= \ \frac{2}{6} \ = \ \frac{1}{3}.\end{aligned}$$

Next notice that features like these can be modeled as Normal random variables with the following MLE's of the means and variances:

	Male		Female	
	Mean	Variance	Mean	Variance
Shoe size	43.25	2.9	38	2
Height	177.5	41.7	165	50
Max Speed	6.75	0.42	6.75	0.125

Then we can estimate the marginal conditionals of our new datum $\mathbf{x} = [42 \;\; 180 \;\; 5.5]^\mathsf{T}$ according to the Normal distribution to obtain:

$$\begin{aligned}
\hat{\mathbb{P}}(\mathrm{x}_1|\mathrm{y} = \text{male}) &= \frac{1}{\sqrt{2\pi(2.9)}} e^{-\frac{(42-43.25)^2}{2(2.9)}} = 0.1787 \\
\hat{\mathbb{P}}(\mathrm{x}_2|\mathrm{y} = \text{male}) &= \frac{1}{\sqrt{2\pi(41.7)}} e^{-\frac{(180-177.5)^2}{2(41.7)}} = 0.0573 \\
\hat{\mathbb{P}}(\mathrm{x}_3|\mathrm{y} = \text{male}) &= \frac{1}{\sqrt{2\pi(0.42)}} e^{-\frac{(5.5-6.75)^2}{2(0.42)}} = 0.0948.
\end{aligned}$$

With this, we can compute the *naive* posterior:

$$\hat{\mathbb{P}}(\mathrm{y} = \text{male}|\mathbf{x}) = \mathbb{P}(y = \text{male}) \prod_{\mathrm{j}=1}^{\mathrm{D}} \hat{\mathbb{P}}(\mathrm{x_i}|\mathrm{y} = \text{male}) = \frac{2}{3}(0.1787)(0.0573)(0.0948) = 6.4745 \times 10^{-4}.$$

Similarly, one can estimate $\hat{\mathbb{P}}(\mathbf{x}|\mathrm{y} = \text{female})$, and choose the class with the highest posterior. What would be your conclusion, is the killer male or female?

10.3 Bayesian Networks

A less *naive* strategy drops the independence assumption, and considers *limited* dependencies between a *query* variable y, *evidence* variables x_j whose values are given, and *hidden* variables x_j whose values are unknown. The dependencies are determined by a directed acyclic graph $\mathcal{G}$, and a set of conditional probability distributions, where each node of the graph represents a variable that is *conditionally independent* on its non-descendants given its parents, which are indicated by the edges of the graph. This way, the joint probability simplifies as:

$$\begin{aligned}
\mathbb{P}(y, x_1, x_2, \ldots, x_\mathrm{D}) &= \mathbb{P}(y) \prod_{\mathrm{j}=1}^{\mathrm{D}} \mathbb{P}(x_\mathrm{j}|y, x_1, x_2, \ldots, x_{\mathrm{j}-1}) \\
&= \mathbb{P}(y) \prod_{\mathrm{j}=1}^{\mathrm{D}} \mathbb{P}(x_\mathrm{j}|\mathcal{P}_\mathcal{G}(x_\mathrm{j})),
\end{aligned}$$

where the first equality is the Law of Total Probability, and $\mathcal{P}_\mathcal{G}(x_\mathrm{j})$ are the parents of node x_j according to the graph $\mathcal{G}$. Bayesian Networks can be used to answer questions like: what is the probability that $y = \mathrm{y}$ given that $x_\mathrm{j} = \mathrm{x_j}$ for every j in a subset of $\{1, \ldots, \mathrm{D}\}$. For example, consider the following related binary random variables:

(B) Burglary occurs at your home,

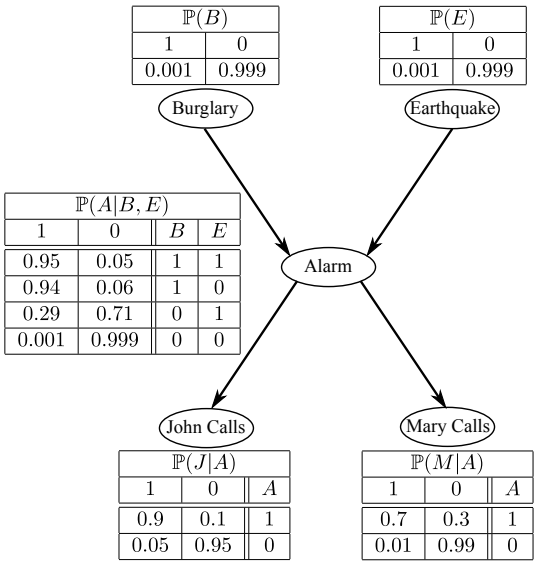
(E) Earthquake occurs at your home,

(A) Alarm goes off,

(J) John calls to report the alarm,

(M) Mary calls to report the alarm,

and suppose a Burglary or Earthquake can trigger an alarm, which may make John or Mary call to report the alarm, according to the following graph and conditional probabilities:



$\mathbb{P}(B)$	
1	0
0.001	0.999

$\mathbb{P}(E)$	
1	0
0.001	0.999

$\mathbb{P}(A\|B,E)$			
1	0	B	E
0.95	0.05	1	1
0.94	0.06	1	0
0.29	0.71	0	1
0.001	0.999	0	0

$\mathbb{P}(J\|A)$		
1	0	A
0.9	0.1	1
0.05	0.95	0

$\mathbb{P}(M\|A)$		
1	0	A
0.7	0.3	1
0.01	0.99	0

Under this setting, the Bayesian Network assumption simplifies the joint probability into:

$$\mathbb{P}(B, E, A, J, M) \;=\; \mathbb{P}(B)\mathbb{P}(E)\mathbb{P}(A|B,E)\mathbb{P}(J|A), \mathbb{P}(M|A).$$

Using Bayes Rule, Bayesian Networks can be used to answer questions like: what is the probability that there is a burglary given that John called to report it? Here B is playing the role of the *query* variable y, $J = 1$ is playing the role of the *evidence*, and E, A, M are hidden. Notice that this choice was completely arbitrary. Bayesian Networks are flexible enough that any variable can be query, evidence, or hidden. Another advantage of this approach is that it reduces the parameter space. For example, a joint distribution with 5 binary random variables has $2^{5-1} = 16$ parameters, whereas the above has only 10. This gap becomes larger as the number of variables increases. In addition, Bayesian Networks capture dependency in an efficient manner, keeping only the relevant dependencies, using a graphical representation that provides insight and interpretability. The challenge, of course, is to infer the graph and conditional probabilities form the data.

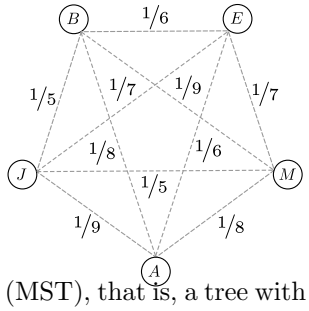
10.3.1 Structure Learning

The first task is to determine the graph structure given a set of training samples. Since Bayesian Networks are flexible enough that any variable can be query, evidence, or hidden, we will assume the training data has the form $\mathbf{x}_1, \ldots, \mathbf{x}_\mathrm{N} \in \mathbb{R}^{\mathrm{D}+1}$, where any of the $\mathrm{D}+1$ features can play the role of the query variable y. Unfortunately, the number of possible graphs is super-exponential in the number of variables, turning the problem of finding the optimal structure NP-complete. Hence, typical heuristics limit the search space of possible structures, like the Chow-Liu algorithm, which infers a Bayesian Network with a tree structure that maximizes the likelihood of the training data using the following steps:

- Estimate the weight between each pair of nodes/variables as their mutual information:

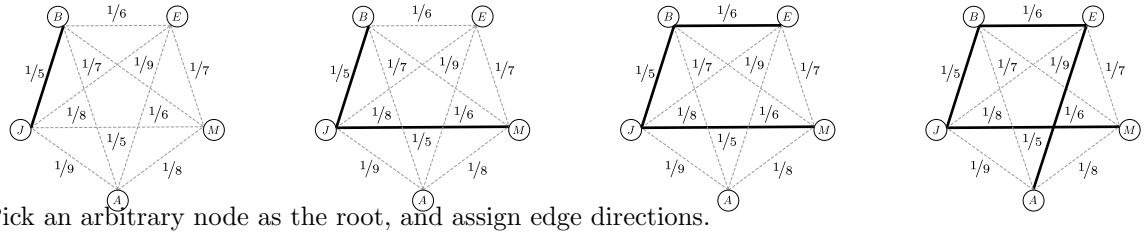
$$I(x_\mathrm{j}, x_\mathrm{k}) \;=\; \sum_{\mathrm{x_j}} \sum_{\mathrm{x_k}} \mathbb{P}(\mathrm{x_j}, \mathrm{x_k}) \log_2 \frac{\mathbb{P}(\mathrm{x_j}, \mathrm{x_k})}{\mathbb{P}(\mathrm{x_j})\mathbb{P}(\mathrm{x_k})}.$$

  In our example, imagine data produces the following weights:

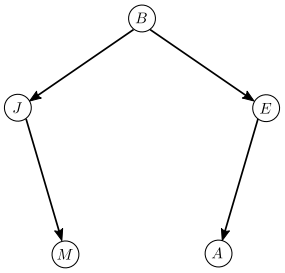


- Find a *maximum-weight spanning tree* (MST), that is, a tree with highest possible weights that connects all nodes in the graph, for example, using Kruskal's algorithm:
  - Start with a graph with $\mathrm{D}+1$ nodes (one per variable), and no edges.
  - Sort all the weights in descending order.
  - For each $\mathrm{t} = 1, 2, \ldots, \mathrm{D}(\mathrm{D}+1)/2$, add the edge corresponding to the t^{th} largest weight unless it creates a cycle.

  In our example the progression as t increases would look like:



- Pick an arbitrary node as the root, and assign edge directions.



The intuition behind the Chow-Liu algorithm is that the mutual information is a proxy of the log-likelihood of the graph $\mathcal{G}$ given i.i.d. data $\mathbf{X} = [\, \mathbf{x}_1 \;\; \mathbf{x}_2 \;\; \cdots \;\; \mathbf{x}_\mathrm{N} \,]^\mathsf{T}$, because

$$\begin{aligned}
\frac{1}{\mathrm{N}} \log \mathbb{P}(\mathbf{X}|\mathcal{G}) \;&=\; \frac{1}{\mathrm{N}} \log \prod_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{P}(\mathbf{x}_\mathrm{i}|\mathcal{G}) \;=\; \frac{1}{\mathrm{N}} \log \prod_{\mathrm{i}=1}^{\mathrm{N}} \prod_{\mathrm{j}=1}^{\mathrm{D}+1} \mathbb{P}\big(\mathrm{x_{ij}}\big|\mathcal{P}_\mathcal{G}(\mathrm{x_{ij}})\big) \\
&=\; \sum_{\mathrm{j}=1}^{\mathrm{D}+1} \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathbb{P}\big(\mathrm{x_{ij}}\big|\mathcal{P}_\mathcal{G}(\mathrm{x_{ij}})\big) \;\xrightarrow[\mathrm{N}\to\infty]{}\; \sum_{\mathrm{j}=1}^{\mathrm{D}+1} \mathbb{E}\left[\log \mathbb{P}\big(x_\mathrm{j}\big|\mathcal{P}_\mathcal{G}(x_\mathrm{j})\big)\right] \\
&=:\; -\sum_{\mathrm{j}=1}^{\mathrm{D}+1} H\big(x_\mathrm{j}\big|\mathcal{P}_\mathcal{G}(x_\mathrm{j})\big) \;=\; \sum_{\mathrm{j}=1}^{\mathrm{D}+1} \left[I\big(x_\mathrm{j}, \mathcal{P}_\mathcal{G}(x_\mathrm{j})\big) - H(x_\mathrm{j})\right],
\end{aligned}$$

where the convergence follows by the Law of Large Numbers, and the last equality is the information theory identity $I(x,y) = H(x) - H(x|y)$; see for example Section 2.4 in *Elements of Information Theory*, by Thomas Cover and Joy Thomas, Second Edition, John Wiley & Sons.

Since $H(x_{\mathrm{j}})$ does not depend on $\mathcal{G}$, the maximum likelihood estimator can be determined through the following optimization

$$\max_{\mathcal{G}} \sum_{\mathrm{j}=1}^{\mathrm{D}+1} I\big(\mathrm{x_j}, \mathcal{P}_{\mathcal{G}}(\mathrm{x_j})\big).$$

Since each node in a tree only has one parent, in the case of trees this can be further simplified into

$$\max_{\mathcal{G}} \sum_{(\mathrm{j},\mathrm{k})\in\mathcal{G}} I(\mathrm{x_j}, \mathrm{x_k}),$$

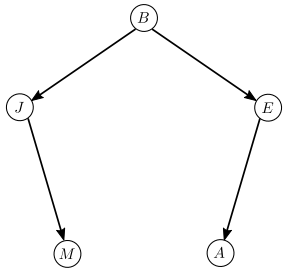
where the notation $(\mathrm{j},\mathrm{k}) \in \mathcal{G}$ indicates that $\mathcal{G}$ has an edge connecting nodes i and k. Finally, the choice of the root does not affected the optimization, because mutual information is symmetric.

10.3.2 Parameter Learning

Given a graph structure $\mathcal{G}$, either because it was known a priori or because it was estimated, the next step is to infer the parameter $\boldsymbol{\theta}$ of the problem, that is, the conditional probabilities $\mathbb{P}(x_{\mathrm{j}}|\mathcal{P}_{\mathcal{G}}(x_{\mathrm{j}}))$. Unsurprisingly, one of the main approaches to do this is maximum likelihood, i.e.,

$$\hat{\boldsymbol{\theta}} \;=\; \arg\max_{\boldsymbol{\theta}} \; \log \mathbb{P}(\mathbf{X}|\mathcal{G},\boldsymbol{\theta}).$$

Continuing with our alarm example, suppose we are given the estimated tree structure above, and the following data:



			Samples									
			1	2	3	4	5	6	7	8	9	10
Variables	x_0	Burglary	1	0	0	0	1	0	0	0	0	1
	x_1	John	1	0	0	0	1	0	0	1	0	0
	x_2	Earthquake	0	0	0	0	1	0	0	0	0	0
	x_3	Marie	0	0	0	0	1	0	0	0	0	1
	x_4	Alarm	1	0	0	0	1	0	0	0	0	1

Our goal reduces to finding the MLE of $\mathbb{P}(x_{\mathrm{j}}|x_{\mathrm{k}})$ for each edge $(\mathrm{j},\mathrm{k}) \in \mathcal{G}$, which is given by the fraction of samples in each case, i.e.,

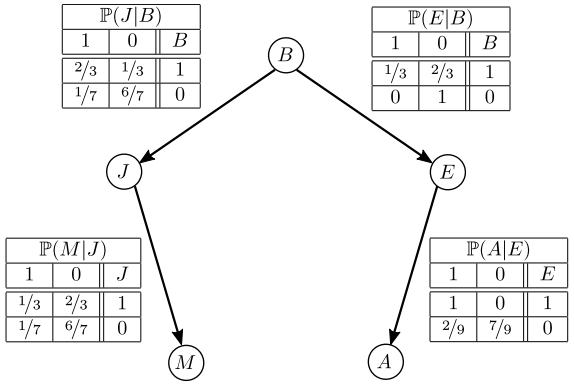
$$\hat{\mathbb{P}}(\mathrm{x_j}|\mathrm{x_k}) \;=\; \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{1}_{\{x_{\mathrm{j}}=\mathrm{x_{ij}}, x_{\mathrm{k}}=\mathrm{x_{ik}}\}}}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{1}_{\{x_{\mathrm{k}}=\mathrm{x_{ik}}\}}}.$$

For instance,

$$\hat{\mathbb{P}}(J=1|B=1) \;=\; \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{1}_{\{J_{\mathrm{i}}=1, B_{\mathrm{i}}=1\}}}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{1}_{\{B_{\mathrm{i}}=1\}}} \;=\; \frac{2}{3},$$

$$\hat{\mathbb{P}}(J=1|B=0) \;=\; \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{1}_{\{J_{\mathrm{i}}=1, B_{\mathrm{i}}=0\}}}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{1}_{\{B_{\mathrm{i}}=0\}}} \;=\; \frac{1}{7}.$$

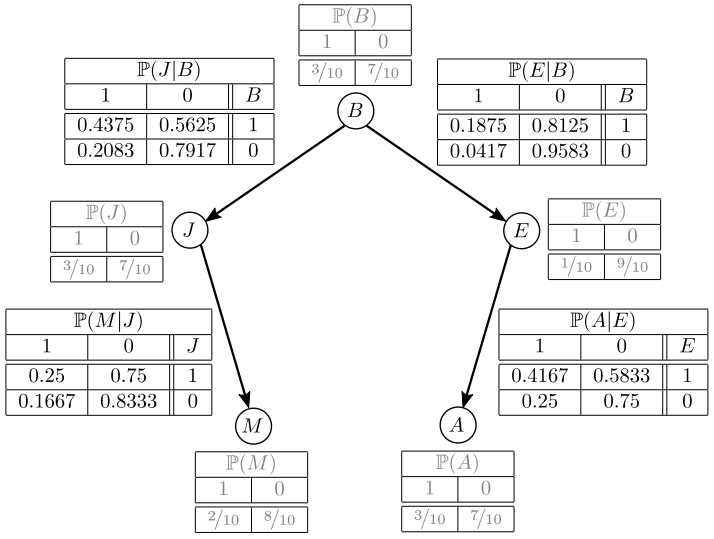
Continuing this way we can obtain all the parameters:



Recall from our discussion above that if a feature x_j has no samples that take value x_j given feature x_k, then $\mathbb{P}(x_j|x_k)$ will be equal to zero, which would also zero out the calculation of the posterior $\mathbb{P}(x_k|x_j)$, neglecting the information in all other probabilities involved. This is exactly what happened with $\mathbb{P}(E|B=0)$ and $\mathbb{P}(A|E=1)$. To avoid this issue we can use Laplace smoothing, Lidstone smoothing, or *m-estimators*, given by:

$$\hat{\mathbb{P}}(x_j|x_k) = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{x_j=x_{ij}, x_k=x_{ik}\}} + m\mathbb{P}(x_j)}{\sum_{i=1}^{N} \mathbb{1}_{\{x_k=x_{ik}\}} + m},$$

where m is a regularization term representing the number of *virtual* samples that we are adding to our dataset, so that no probability is ever estimated to be exactly zero. Choosing $m = 5$ in our example, we would obtain the following m-estimates instead:



Given these estimates, we can answer questions like: what is the probability that there was a Burglary given that only Mary called to report an alarm, and that there was no earthquake:

$$\mathbb{P}(B|J,E,M,A) = \frac{\mathbb{P}(B)\mathbb{P}(J,E,M,A|B)}{\mathbb{P}(J,E,M,A)} = \frac{\mathbb{P}(B)\mathbb{P}(A|E)\mathbb{P}(M|J)\mathbb{P}(E|B)\mathbb{P}(J|B)}{\mathbb{P}(J,E,M,A)}$$

where the first equality follows by Bayes Rule, and the second one by the definition of conditional probability. What probability do you obtain? Would you conclude that there was a Burglary, or not?

10.4 Maximum A Posteriori (MAP) Estimators

Like Laplace, and M-estimators, and the MLE, there exists another family of estimators, called *maximum a posteriori* (MAP), which uses a Bayesian approach. Given data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\mathsf{T}$ where each $\mathbf{x}_i$ is independently and identically distributed according to $\mathbb{P}(\boldsymbol{x}|\boldsymbol{\theta})$, the goal is to find the parameter $\boldsymbol{\theta}$ that maximizes the posterior distribution, i.e.,

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} \;:=\; \arg\max_{\boldsymbol{\theta}} \; \mathbb{P}(\boldsymbol{\theta}|\mathbf{X}).$$

Using Bayes Rule we can rewrite this as:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} \;=\; \arg\max_{\boldsymbol{\theta}} \; \frac{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathbf{X})} \;=\; \arg\max_{\boldsymbol{\theta}} \; \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta}). \tag{10.7}$$

Notice that this is quite similar as the (MLE):

$$\hat{\boldsymbol{\theta}}_{\text{ML}} \;:=\; \arg\max_{\boldsymbol{\theta}} \; \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}),$$

except for the $\mathbb{P}(\boldsymbol{\theta})$ factor, which accounts for the *prior* probability of each parameter. Intuitively, the prior leans our estimator towards our educated guess of what it might be. We formalize this in the next proposition.

Proposition 10.1. If there is no prior information, the MAP is equal to the MLE.

Proof. Having *no* prior information is equivalent to having $\boldsymbol{\theta}$ being uniformly distributed. In this case $\mathbb{P}(\boldsymbol{\theta})$ is a constant. Hence

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} \;:=\; \arg\max_{\boldsymbol{\theta}} \; \mathbb{P}(\boldsymbol{\theta}|\mathbf{X}) \;=\; \arg\max_{\boldsymbol{\theta}} \; \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta}) \;=\; \arg\max_{\boldsymbol{\theta}} \; \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) \;=:\; \hat{\boldsymbol{\theta}}_{\text{ML}}.$$

□

Remark 10.1. Notice that the Naive Bayes estimator $\hat{y}$ in (10.5) is one particular instance of (10.7) under the *naive* assumption that each $\mathrm{x_j}$ is independent.

Example 10.6 (Pharmaceutics)**.** Scientists at a big pharmaceutical company have designed a COVID-19 vaccine, and want to estimate its probability of success $\mathrm{p}^\star$. To this end they will conduct a clinical trial where they will test their treatment on N individuals, and record whether they react favorably. This can be modeled as

$$x_1, \ldots, x_N \overset{iid}{\sim} \text{Bernoulli}(\mathrm{p}^\star),$$

and the goal is to estimate $\mathrm{p}^\star$.

Pharmaceuticals design many treatments. Testing them on humans is difficult and expensive. Hence they first experiment in-vitro or with animals to find the most effective ones. The particular treatment that we are studying has already been tested in vitro, mice, rabbits and chimpanzees, and has proven to be very effective. Hence we expect *a priori* that $\mathrm{p}^\star$ will be closer to 1 than to 0. Thus, a good model

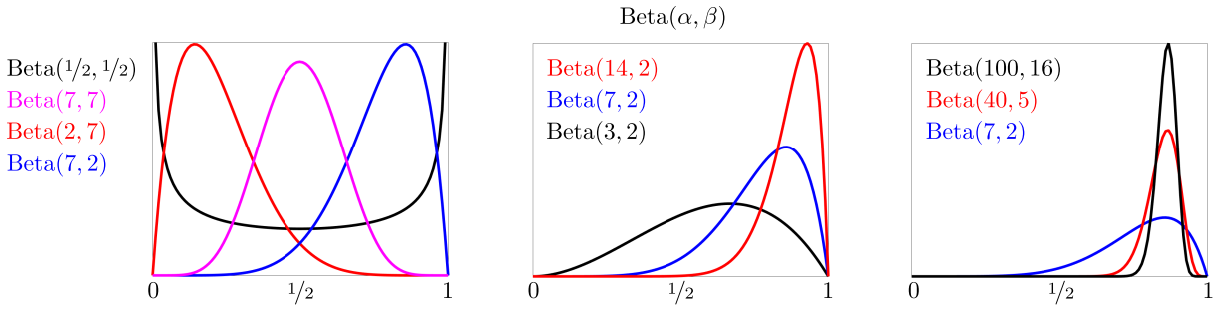


Figure 10.1: Beta(α, β) densities are good models for prior distributions of proportions (like the probability of success $\mathrm{p}^\star$). Loosely speaking, the gap between α and β determines where we a priori *think* $\mathrm{p}^\star$ is; the magnitudes of α and β determine how *confident* we are. This way the parameters α and β determine our a priori *bias* and *certainty*. In our Pharmaceutics example, if we believe the probability of success $\mathrm{p}^\star$ to be closer to 1, we can model it as Beta(α, β) with $\alpha > \beta > 1$, so that $\mathbb{P}(p)$ is *biased* towards 1. If we are *somewhat* certain, we can take $\alpha = 7$ and $\beta = 2$. If we are *extremely* certain, we can choose α and/or β to be much larger.

for the prior $\mathbb{P}(p)$ would be the Beta density

$$\mathbb{P}(p) \;=\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, p^{\alpha-1}(1-p)^{\beta-1}$$

with parameters $\alpha > \beta > 1$, so that the density is skewed towards 1 (see Figure 10.1 for some intuition). Using this prior information, the pharmaceutical will use a bayesian approach to estimate $\mathrm{p}^\star$. The likelihood function (conditional) of our data $\mathbf{X} = [\,\mathrm{x}_1 \;\; \mathrm{x}_2 \;\; \cdots \;\; \mathrm{x}_\mathrm{N}]^\mathsf{T}$ is

$$\mathbb{P}(\mathbf{X}|p) \;=\; p^{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{i}}\,(1-p)^{\mathrm{N}-\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{i}} \;=\; p^{\mathbf{1}^\mathsf{T}\mathbf{X}}\,(1-p)^{\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{X}},$$

where $\mathbf{1} \in \mathbb{R}^\mathrm{N}$ is the vector with all ones, so that $\mathbf{1}^\mathsf{T}\mathbf{X}$ is shorthand for $\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{i}$. Hence

$$\mathbb{P}(p|\mathbf{X}) \propto \mathbb{P}(\mathbf{X}|p)\mathbb{P}(p) = \left(p^{\mathbf{1}^\mathsf{T}\mathbf{X}}\,(1-p)^{\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{X}}\right)\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, p^{\alpha-1}(1-p)^{\beta-1}\right) \propto p^{\mathbf{1}^\mathsf{T}\mathbf{X}+\alpha-1}\,(1-p)^{\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{X}+\beta-1},$$

and so we recognize $\mathbb{P}(p|\mathbf{X})$ to be the Beta(α', β') density with parameters $\alpha' = \mathbf{1}^\mathsf{T}\mathbf{X} + \alpha$ and $\beta' = \mathrm{N} - \mathbf{1}^\mathsf{T}\mathbf{X} + \beta$ (here we are omitting the normalization factor $\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')}$, which we know $\mathbb{P}(p|\mathbf{X})$ must have, because it is a density and must integrate to 1). It follows that $\hat{p}_\mathrm{MAP} = \arg\max_p \mathbb{P}(p|\mathbf{X})$ is the point that maximizes the Beta(α', β') density, i.e., its mode: if $\alpha', \beta' > 1$, this is given by $\frac{\alpha'-1}{\alpha'+\beta'-2}$; if α' or $\beta' < 1$, it is one of the extreme points $\{0, 1\}$; if $\alpha' = \beta' = 1$, then Beta(α', β') = Uniform$[0, 1]$.

Definition 10.3 (Conjugate prior)**.** Whenever $\mathbb{P}(\boldsymbol{\theta})$ has the same form (but possibly different parameters) as $\mathbb{P}(\boldsymbol{\theta}|\mathbf{X})$, we say that $\mathbb{P}(\boldsymbol{\theta})$ is a *conjugate prior* of $\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})$.

Example 10.7. From Example 10.6 we conclude that *Beta*(α, β) is the conjugate prior of *Bernoulli*(p).

10.5 The Bernstein-von Mises Theorem

Notice that to find the MAP one inherently needs to know the posterior distribution $\mathbb{P}(\boldsymbol{\theta}|\mathbf{X})$. For instance, in Example 10.6 we first discovered that the posterior distribution was Beta, and then found the parameter that maximized such distribution (in this case, the mode). In general, finding the posterior may not always be as easy and clean as in Example 10.6. Fortunately, the Bernstein-von Mises Theorem shows that for a sufficiently large number of samples, the posterior distribution is Normal, centered around the true parameter with diminishing variance, independent of the prior.

Definition 10.4 (Total Variation Distance)**.** Given two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same sigma-algebra $\mathcal{A}$, their *total variation distance* is defined as

$$\|\mathbb{P}-\mathbb{Q}\|_{\text{TV}} \;:=\; \sup_{E\in\mathcal{A}} \big|\mathbb{P}(E)-\mathbb{Q}(E)\big|.$$

Intuitively, it is the largest possible difference that the two distributions assign to the same event.

Theorem 10.1 (Bernstein-von Mises)**.** Let $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_{\text{N}}]^{\mathsf{T}}$, with $\mathbf{x}_{\text{i}} \overset{iid}{\sim} \mathbb{P}(\boldsymbol{x}|\boldsymbol{\theta})$. Suppose $\dfrac{\text{d}^2 \log \mathbb{P}(\boldsymbol{x}|\boldsymbol{\theta})}{\text{d}\boldsymbol{\theta}^2}$ exists, and that the prior $\mathbb{P}(\boldsymbol{\theta})$ is continuous and strictly positive on the true parameter $\boldsymbol{\theta}^\star$. Also suppose that for every $\epsilon>0$ there exists a sequence of tests ϕ_{N} such that

$$\lim_{\text{N}\to\infty} \mathbb{E}_{\boldsymbol{\theta}^\star}[\phi_{\text{N}}] = 0 \qquad \textit{and} \qquad \lim_{\text{N}\to\infty} \sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^\star\|\geq\epsilon} \mathbb{E}_{\boldsymbol{\theta}}[1-\phi_{\text{N}}] = 0.$$

Then as $\text{N}\to\infty$

$$\left\| \mathbb{P}\left(\boldsymbol{\theta}|\mathbf{X}\right) - \mathcal{N}\left(\boldsymbol{\theta}^\star + \frac{1}{\text{N}}\Delta_{\boldsymbol{\theta}^\star}\;,\; \frac{1}{\text{N}}\mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}\right)\right\|_{\text{TV}} \longrightarrow 0,$$

where

$$\Delta_{\boldsymbol{\theta}^\star} \;:=\; \mathbf{I}_{\boldsymbol{\theta}^\star}^{-1} \sum_{\text{i}=1}^{\text{N}} \frac{\text{d}\log\mathbb{P}(\mathbf{x}_{\text{i}}|\boldsymbol{\theta})}{\text{d}\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}, \qquad \textit{and} \qquad \mathbf{I}_{\boldsymbol{\theta}^\star} \;:=\; -\mathbb{E}\left[\frac{\text{d}^2\log\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\text{d}\boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star}\right]$$

denotes the *Fisher-information matrix*.

Proof. The proof follows as a consequence of the central limit theorem, which says that if $z_1,\ldots,z_{\text{N}}$ are i.i.d. with mean μ and variance σ^2, then $\frac{\sqrt{\text{N}}}{\sigma}\left(\sum_{\text{i}=1}^{\text{N}} z_{\text{i}} - \mu\right)$ is asymptotically distributed $\mathcal{N}(0,1)$. For a detailed proof see Section 5.2.2 in *Statistiques Asymptotiques*, by Elisabeth Gassiat. □

The assumptions of the Theorem are quite weak, and satisfied in most common cases. To learn more see pages 145-146 of *Asymptotic Statistics* by Aad van der Vaart, Cambridge University Press, 1998.

Since total variation convergence implies convergence in distribution, Theorem 10.1 implies (with a slight

abuse of notation) that as $\mathrm{N} \to \infty$.

$$\boldsymbol{\theta}|\mathbf{X} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\theta}^\star + \frac{1}{\mathrm{N}}\Delta_{\boldsymbol{\theta}^\star} \ , \ \frac{1}{\mathrm{N}}\mathbf{I}_{\boldsymbol{\theta}^\star}^{-1}\right).$$

Notice the similarity of this result with the asymptotic distribution of the MLE:

The Bernstein-von Mises Theorem establishes an important link between Bayesian inference and Frequentist statistics, showing that the the effect of the prior decreases with the number of samples. To build some intuition, suppose you initially believe that a coin is loaded to favor heads (Bernoulli prior with $p > 1/2$). If you toss that coin a million times and realize that 90% of the time it falls tails, would you still believe the coin favors heads? In other words, would you still believe a posteriori that $p > 1/2$? How many coin tosses would suffice to convince you of the truth? What if your initial belief had been accurate? The Bernstein-von Mises Theorem formalizes this intuition that overwhelming evidence (large samples) will reveal the truth, regardless of a correct or incorrect initial belief (prior). Finally, another consequence of the Bernstein-von Mises Theorem is that MAP converges to the MLE as the number of samples grows.

Example 10.8. In Example 10.6, the MAP is the mode of the posterior distribution $\mathbb{P}(p|\mathbf{X}) = \text{Beta}(\alpha', \beta')$. If $\alpha', \beta' > 1$, this is given by:

$$\hat{p}_{\text{MAP}} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{\mathbf{1}^\mathsf{T}\mathbf{X} + \alpha - 1}{\cancel{\mathbf{1}^\mathsf{T}\mathbf{X}} + \alpha + \mathrm{N} - \cancel{\mathbf{1}^\mathsf{T}\mathbf{X}} + \beta - 2} = \frac{\mathbf{1}^\mathsf{T}\mathbf{X} + \alpha - 1}{\mathrm{N} + \alpha + \beta - 2} \xrightarrow[\mathrm{N}\to\infty]{} \mathrm{p}^\star.$$

where the last implication follows because

$$\frac{\mathbf{1}^\mathsf{T}\mathbf{X}}{\mathrm{N}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i} \xrightarrow[\mathrm{N}\to\infty]{} \mathrm{p}^\star$$

by the Law of Large Numbers. On the other hand, the MLE is the maximizer of the likelihood $\mathbb{P}(\mathbf{X}|p) = \text{Bernoulli}(p)$, given by:

$$\hat{p}_{\text{ML}} = \arg\max_{p\in[0,1]} \mathbb{P}(\mathbf{X}|p) = \arg\max_{p\in[0,1]} p^{\mathbf{1}^\mathsf{T}\mathbf{X}}(1-p)^{\mathrm{N}-\mathbf{1}^\mathsf{T}\mathbf{X}} = \frac{\mathbf{1}^\mathsf{T}\mathbf{X}}{\mathrm{N}} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i} \xrightarrow[\mathrm{N}\to\infty]{} \mathrm{p}^\star.$$

On the other hand, if $\mathrm{p}^\star = 0$ or 1, then so will be $\hat{p}_{\text{MAP}}$ and $\hat{p}_{\text{ML}}$. We thus conclude that $\hat{p}_{\text{MAP}} \to \hat{p}_{\text{ML}}$ as $\mathrm{N} \to \infty$, as implied by Theorem 10.1.

10.6 Simulations

Simulations are not only good to verify our results. They are also good to build intuition, test theories and draw conclusions. In this section we will further study Example 10.6 to compare the MAP and the MLE. We will verify our intuition that if our prior is accurate, the MAP will be better, but if our prior is inaccurate, the MAP will be worse.

Recall that the treatment in Example 10.6 has already shown great results on other organisms, so we believe its probability of success on humans $\mathrm{p}^\star$ to be closer to 1 than to 0. With this in mind we will use a $\text{Beta}(7, 2)$ as prior, so that the density is skewed towards 1. (see Figure 10.1 to build some intuition).

Next we will generate a random vector $\mathbf{X} \in \mathbb{R}^\mathrm{N}$ with i.i.d. Bernoulli($\mathrm{p}^\star$) entries, so that the i^{th} entry in $\mathbf{X}$ simulates whether the i^{th} patient reacted favorably to the treatment. We showed in Example 10.6 that the posterior distribution $\mathbb{P}(p|\mathbf{X})$ is $\text{Beta}(\alpha', \beta')$ with parameters $\alpha' = \mathbf{1}^\mathsf{T}\mathbf{X} + \alpha$ and $\beta' = \mathrm{N} - \mathbf{1}^\mathsf{T}\mathbf{X} + \beta$.

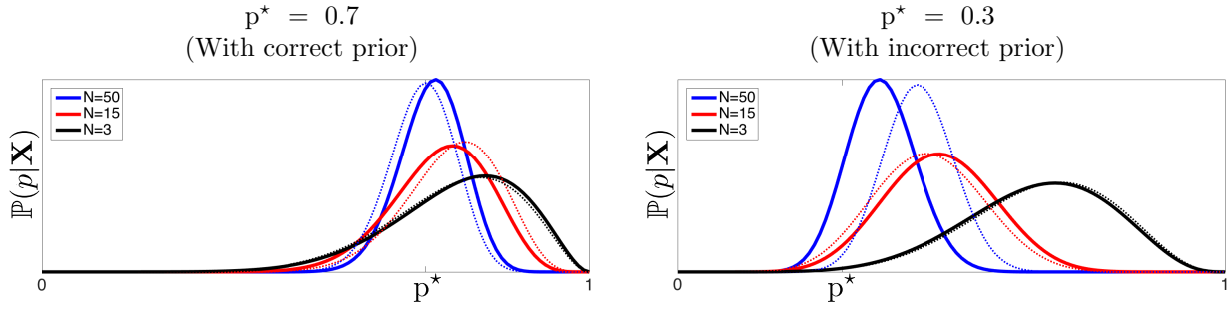


Figure 10.2: Posterior distribution $\mathbb{P}(p|\mathbf{x})$ in Example 10.6. In expectation, $\mathbf{1}^\mathsf{T}\mathbf{x} = \mathrm{N}\mathrm{p}^\star$, so the expected (theoretical) posterior (plotted in solid lines) is $\mathrm{Beta}(\mathrm{N}\mathrm{p}^\star + \alpha, \mathrm{N}(1-\mathrm{p}^\star) + \beta)$. Dotted lines are the posterior distributions given a particular sample. The code for this simulation is in Appendix A.

Let us now consider two scenarios:

(i) $\mathrm{p}^\star = 0.7$. This would be a case when our prior is correct (Figure 10.2—left). We can see that even with a few samples ($\mathrm{N} = 3$), our estimate $\hat{\mathrm{p}}_{\mathrm{MAP}} = \arg\max_{\mathrm{p}\in[0,1]} \mathbb{P}(\mathrm{p}|\mathbf{x})$ would be very close to $\mathrm{p}^\star$. The code for this simulation is in Appendix A.

(ii) $\mathrm{p}^\star = 0.3$. This would be a case when our prior is incorrect (Figure 10.2—right). We can see that unless we have a lot of samples (N large), our estimate $\hat{\mathrm{p}}_{\mathrm{MAP}}$ could be very far from $\mathrm{p}^\star$! In words, the prior is *pulling* the posterior towards it. This is one of the dangers of bayesian estimation: the bias induced by the prior might make it harder to see the truth. What do you think would happen if we use a *stronger* prior, like $\mathbb{P}(\mathrm{p}) = \mathrm{Beta}(100, 16)$? How would this affect the posterior $\mathbb{P}(\mathrm{p}|\mathbf{x})$? Try it out and see; you only need to change a few lines of code. Do the results match your intuition?

Case (ii) shows one of the risks of bayesian estimation. Now the question is: is it worth it? In other words, *if* our prior is correct, do we really have that much to gain? Let us find out by comparing the MAP with the MLE. Since $x_i \overset{iid}{\sim} \mathrm{Bernoulli}(\mathrm{p}^\star)$, it follows that $\hat{p}_{\mathrm{ML}} = \sum_{\mathrm{i}=1}^{\mathrm{N}} x_i$ has a $\mathrm{Binomial}(\mathrm{N}, \mathrm{p}^\star)$ distribution (scaled by $1/\mathrm{N}$). Figure 10.3 shows a comparison of the distributions of $\hat{p}_{\mathrm{MAP}}$ and $\hat{p}_{\mathrm{ML}}$ for different scenarios.

These experiments show that the prior is essentially *biasing* us towards our *beliefs*. If our beliefs are correct, the MAP will be more accurate than the MLE (given the same number of samples N). In contrast, if our beliefs are incorrect, it will take more samples to *correct* the prior, and so the MAP will be more inaccurate. This experiment also verifies that the MAP converges to the MLE as N grows, as implied by Theorem 10.1.

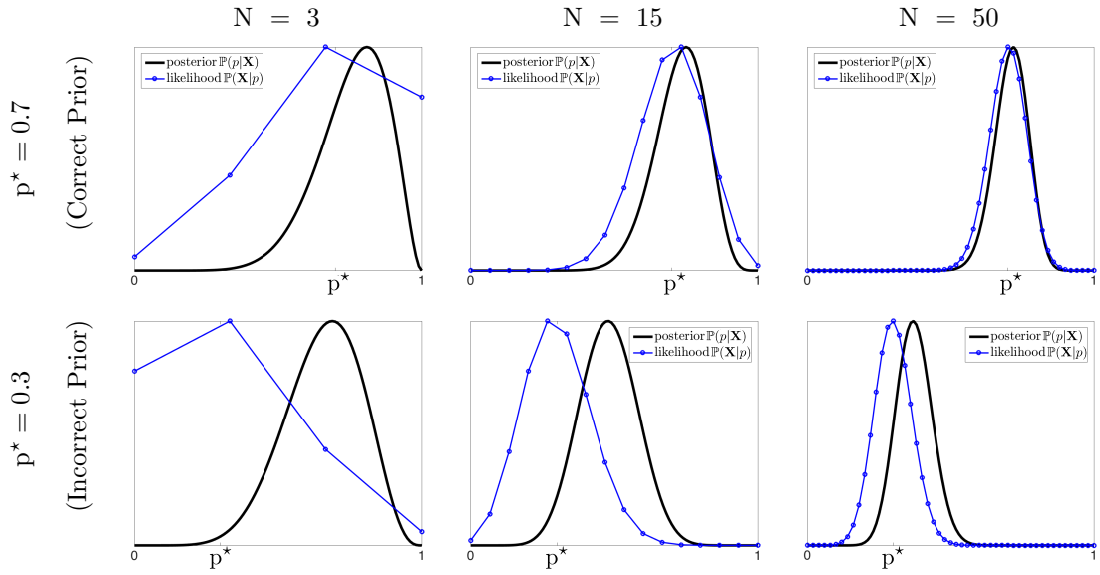


Figure 10.3: Distribution of the MAP and the MLE for different values of $\mathrm{p}^\star$ and N. $\hat{p}_{\mathrm{MAP}} \sim \mathrm{Beta}(\mathrm{N}\mathrm{p}^\star + \alpha, \mathrm{N}(1-\mathrm{p}^\star) + \beta)$ and $\hat{p}_{\mathrm{ML}} \sim \mathrm{Binomial}(\mathrm{N}, \mathrm{p}^\star)$. The code for this is in Appendix B.

A Code for Simulation of Example 10.6

```
clear all; close all; clc; warning('off','all');

% === Code to simulate the posterior distribution of p ===
% === See Example 10.5.

p_star = 0.7;           % True probability of success.
alpha = 7;              % Parameter of prior distribution.
beta = 2;               % Parameter of prior distribution.
NN = [50,15,3];         % Sample sizes we will try.

% Create figure.
figure(1);
axes('Box','on');
hold on;

% Plot expected posterior distribution.
p = 0:0.01:1;           % all possible values of p.
color = ['b','r','k'];  % For plotting.
for n=1:length(NN)

    N = NN(n);                          % Number of samples.
    alpha_prime = N*p_star + alpha;   % Parameter of expected posterior distribution.
    beta_prime = N*(1-p_star) + beta; % Parameter of expected posterior distribution.
    posterior = betapdf(p,alpha_prime,beta_prime);   % Expected posterior distribution.
    plot(p,posterior,color(n),'LineWidth',4);

end

% Legends.
legend(['N=',num2str(NN(1))],['N=',num2str(NN(2))],['N=',num2str(NN(3))],...
    'Interpreter','latex','fontsize',20,'Location','Northwest');

```

```
% Plot posterior distributions for a particular sample [X_1,...X_N].
for n=1:length(NN)

    N = NN(n);                          % Number of samples.
    X = rand([N,1]) < p_star;          % Sample.
    alpha_prime = sum(X) + alpha;      % Parameter of posterior distribution based on sample.
    beta_prime = N - sum(X) + beta;    % Parameter of posterior distribution based on sample.
    posterior = betapdf(p,alpha_prime,beta_prime);    % Posterior distribution based on sample.
    plot(p,posterior,[color(n),':'],'LineWidth',2);

end

% Make figure look sexy.
axis tight;
ylabel('$P(p|\textbf{X})$','Interpreter','latex','fontsize',20);
xlabel('','Interpreter','latex','fontsize',20);
set(gca,'XTick',[0,p_star,1],'xticklabel',{'0','p*','1'},'fontsize',20);
set(gca,'YTick',[],'yticklabel',[]);
set(gcf,'PaperUnits','centimeters','PaperSize',[30,10],'PaperPosition',[0,0,30,10]);

% Save figure.
set(gcf, 'renderer','default');
figurename = 'MAP.pdf';
saveas(gcf,figurename);
```

B Code for Comparison of MAP and MLE in Figure 10.3

```
clear all; close all; clc; warning('off','all');

% === Code to simulate the posterior distribution of p ===
% === See Example 10.5

p_star = 0.7;           % True probability of success.
alpha = 7;              % Parameter of prior distribution.
beta = 2;               % Parameter of prior distribution.
NN = [50,15,3];         % Sample sizes we will try.

% Plot distributions of the MAP and the MLE.
for n=1:length(NN)

    N = NN(n);  % Number of samples.

    % Create figure
    figure(n);
    axes('Box','on');
    hold on;

    % MAP distribution.
    p = 0:0.01:1;                       % all possible values of p (continuous).
    alpha_prime = N*p_star + alpha;     % Parameter of expected posterior distribution.
    beta_prime = N*(1-p_star) + beta; % Parameter of expected posterior distribution.
    posterior = betapdf(p,alpha_prime,beta_prime);    % Expected posterior distribution.
    h1 = plot(p,posterior,'k','LineWidth',4);

    % MLE distribution.
    p = 0:N;                                % all possible values of rho (discrete).
    likelihood = binopdf(p,N,p_star);     % Distribution of the MLE.
    likelihood = likelihood/max(likelihood)*max(posterior);
    h2 = plot(p/N,likelihood,'b-o','LineWidth',2);
```

```
33
34      % Make figure look sexy.
35      axis tight;
36      ylabel('','Interpreter','latex','fontsize',20);
37      xlabel('','Interpreter','latex','fontsize',20);
38      set(gca,'XTick',[0,p_star,1],'xticklabel',{'0','p*','1'},'fontsize',20);
39      set(gca,'YTick',[],'yticklabel',[]);
40      title(['N$ = ',num2str(N),'$'],'Interpreter','latex','fontsize',25);
41      legend({'posterior $P(p|\textbf{X})$',...
42          'likelihood $P(\textbf{X}|p)$'},...
43          'Interpreter','latex','fontsize',20,'Location','Northwest');
44      set(gcf,'PaperUnits','centimeters','PaperSize',[20,15],'PaperPosition',[0,0,20,15]);
45
46      % Save figure.
47      set(gcf, 'renderer','default');
48      figurename = ['MAPvsMLE_',num2str(N),'.pdf'];
49      saveas(gcf,figurename);
50
51  end
```