Repetion from lecture: model for a coin

Recommended reading / reference: Murphy 3.1, 3.2, 3.3.

In the lecture part on MLE, we discussed **learning a model for an unknown coin** on the basis of N iid tosses of that coin. Each toss F_i is $Ber(F_i|\Theta)$ distributed: $F_i \sim Ber(\Theta)$. Let a sequence of the N coin tosses (the data) (e.g. (H,T,H,T,H,H,H,T,H,H,H)) be represented by $\mathcal{D}$ (formally: $\mathcal{D} := \mathcal{D}_N$ is a (vector) random variable mapping $\Omega = \{H,T\}^N$ to $\{0,1\}^N$). The likelihood is:

$$
\begin{aligned}
p(D|\Theta) &= \prod_{i=1}^{N} \Theta^{\mathbb{1}(F_i=H)}(1-\Theta)^{\mathbb{1}(F_i=T)} &(1)\\
&= \Theta^{N_H}(1-\Theta)^{N-N_H} &(2)
\end{aligned}
$$

A simple MLE point estimation for Θ via $d/d\Theta\ p(\mathcal{D}|\Theta) \overset{!}{=} 0$ gives

$$argmax_{\Theta} p(D|\Theta) = \frac{N_H}{N}$$

we know that for small $\mathcal{D}$ MLE is error-prone, so we use a conjugate prior $p(\Theta|a,b) \sim Beta(\Theta|a,b)$. We then get for the MAP (posterior $\propto$ likelihood * prior):

$$
\begin{aligned}
p(\Theta|D) &\propto p(D|\Theta)p(\Theta|a,b) &(3)\\
&\propto \Theta^{N_H}(1-\Theta)^{N-N_H} Beta(\Theta|a,b) &(4)\\
&\propto Beta(\Theta|N_H+a, N-N_H+b) &(5)
\end{aligned}
$$

So because $Beta$ is normalized we conclude that

$$p(\Theta|D) = Beta(\Theta|N_H+a, N-N_H+b)$$

Thus a MAP point estimation of Θ via $d/d\Theta\ p(\Theta|\mathcal{D}) \overset{!}{=} 0$ (in fact nothing but the mode (the extremal point) of the Beta distribution) gives

$$argmax_{\Theta} p(\Theta|D) = \frac{N_H+a-1}{N+a+b-2}$$

For a posterior prediction of a new coin toss F, one could either use

- the Θ won from MLE leading to a posterior predictive distribution

$$p(F|\Theta_{MLE}) = \Theta_{MLE}^{\mathbb{1}(F=H)}(1-\Theta_{MLE})^{\mathbb{1}(F=T)}$$

- or one could use the Θ_{MAP} won from MAP leading to a posterior predictive distribution

$$p(F|\Theta_{MAP}) = \Theta_{MAP}^{\mathbb{1}(F=H)}(1-\Theta_{MAP})^{\mathbb{1}(F=T)}$$

- OR one could follow a full Bayesian approach to deriving the posterior predictive distribution by not using a point estimation for Θ but rather integrating over all possible Θ:

$$\begin{aligned}
p(F=H|\mathcal{D}) &= \int_0^1 p(F=H|\Theta)p(\Theta|D)d\Theta & (6)\\
&= \int_0^1 Ber(F=H|\Theta)Beta(\Theta|N_H+a, N-N_H+b)d\Theta & (7)\\
&= \int_0^1 \Theta Beta(\Theta|N_H+a, N-N_H+b)d\Theta & (8)\\
&= E_{Beta(\Theta|N_H+a, N-N_H+b)}[\Theta] & (9)\\
&= \frac{N_H+a}{N+a+b} & (10)
\end{aligned}$$

So in essence when switching from MAP to full Bayesian, we replace the mode (the argmax) of the Beta with the mean. So even for a uniform prior ($a=1$, $b=1$), the full Bayesian approach adds 1 to the numerator and denominator. This is referred to as 'Laplace rule'.

The first line can be used, because $F \perp\!\!\!\perp \mathcal{D}|\Theta$ (once we know the coin (once we know Θ) the new coin toss is independent of the data). The first line (for the discrete case) has been proved in the tutor exercises of tutorial 2.
The = in the second line follows, because the Beta distribution is, of course, normalized to one, so we don't need to worry about the normalizing constant in the $\propto$ calculations above.

Problem 1: using a different data representation

What do we need to change in the upper calculations if

- we switch to another representation of the data $\mathcal{D} := \mathcal{D}_N := X_N : \Omega = \{H,T\}^N \to \mathbb{N}$ (where $X_N = N_H$ means N_H heads have occurred in N tosses) and
- we are interested in predicting the number of heads $X := X_{N+1}$ after a new toss?

Solution to problem 1

We don't have to change anything fundamental, since (N_H, N) is a sufficient statistics of the problem for both cases. equation 2 would change only by multiplication with the constant term $\binom{N}{N_H}$

$$p(\mathcal{D}|\Theta) = p(X_N|\Theta) = \binom{N}{N_H}\Theta^{N_H}(1-\Theta)^{N-N_H}$$

so that the MLE estimation for Θ is unchanged.

For the MAP estimate there is also no relevant change:

$$\begin{aligned} p(\Theta|D) = p(\Theta|X_N) = \propto p(X_N|\Theta)p(\Theta|a,b) && (11)\\ \propto \binom{N}{N_H}\Theta^{N_H}(1-\Theta)^{N-N_H}Beta(\Theta|a,b) && (12)\\ \propto Beta(\Theta|N_H+a, N-N_H+b) && (13) \end{aligned}$$

so that the MAP estimation for Θ is also unchanged.

However for the posterior predictive distributions using Θ_{MLE} and Θ_{MAP} and the full Bayesian approach, we have to revert to the posterior predictive distribution $p(F|\Theta)$ for F (the next coin toss) and use this immediately to calculate $p(X_{N+1}|\Theta)$: If the known data is $X_N = N_H$, we have for the MLE and MAP cases:

$$p(X_{N+1} = x|\Theta) = \begin{cases} p(F = H|\Theta) & \text{if } x = N_H + 1 \\ p(F = T|\Theta) & \text{if } x = N_H \\ 0 & \text{else} \end{cases} \tag{14}$$

For the full Bayesian case we **cannot** naively use something like

$$\begin{aligned} p(X_{N+1}|X_N) &= \int_0^1 p(X_{N+1}|\Theta)p(\Theta|X_N)d\Theta \\ &\propto \int_0^1 Bin(X_{N+1}|\Theta, N+1)Beta(\Theta|X_N+a, N-X_N+b)d\Theta \\ &= \dots \end{aligned}$$

because $X_{N+1} \not\perp\!\!\!\perp X_N|\Theta$. (The number of heads after $N+1$ tosses is not conditionally independent from the number of heads after N tosses, (even) if we know the model for the coin (know Θ)).

We must rather calculate

$$\begin{aligned} p(X_{N+1} = N_H + 1|\mathcal{D}) &= p(X_{N+1} = N_H + 1|X_N = N_H) && (15)\\ &= p(F = H|X_N = N_H) && (16)\\ &= \int_0^1 p(F = H|\Theta)p(\Theta|X_N = N_H)d\Theta && (17)\\ &= \int_0^1 Ber(F = H|\Theta)Beta(\Theta|N_H + a, N - N_H + b)d\Theta && (18)\\ &= \int_0^1 \Theta Beta(\Theta|N_H + a, N - N_H + b)d\Theta && (19)\\ &= E_{Beta(\Theta|N_H+a,N-N_H+b)}[\Theta] && (20)\\ &= \frac{N_H + a}{N + a + b} && (21) \end{aligned}$$

(second to third line because $F \perp\!\!\!\perp X_N|\Theta$) which is no change as well.

Problem 2: Dirichlet-multinomial model

Generalize the model for a two sided coin from the lecture to a K-sided dice! Compute the MLE and MAP point estimations for the model parameters! What is the posterior predictive distribution for MLE, MAP, and the full Bayesian approach?

Solution to problem 2

Recommended reading / reference: Murphy 3.4

Observing N dice rolls, we have $\mathcal{D} = (x_1, x_2, \dots, x_N)$ with $x_i \in \{1, 2, \dots; K\}$. Assuming iid, we have for the likelihood

$$p(\mathcal{D}|\Theta) = \prod_{k=1}^{K} \Theta_k^{N_k} \propto Mu(N_1, \dots, N_K|\Theta_1, \dots, \Theta_k)$$

with $\sum_k \Theta_k = 1$, and N_k being the number of occurrences of side k. We thus have K parameters Θ_k or in a more compact notation a $K-$vector Θ to learn for our parametric model. $Mu(N_1, \dots, N_K|\Theta_1, \dots, \Theta_k)$ is the Multinomial distribution $Mu(N_1, \dots, N_K|\Theta_1, \dots, \Theta_k) = N!/(N_1!N_2! \dots N_k!) \prod_{k=1}^{K} \Theta_k^{N_k}$

For computing the posterior, we must find a conjugate (matching) prior. In the coin case the likelihood was Ber or $Beta$ distributed, so we chose a $Beta$ prior. Here, we should use a generalization of the Beta distribution, the Dirichlet distribution. So we use

$$p(\Theta|\alpha) \sim Dir(\Theta|\alpha) \propto \prod_{k=1}^{K} \Theta_k^{\alpha_k - 1}$$

as a prior.

doing so results in the posterior

$$\begin{aligned}
p(\Theta|D) &\propto p(D|\Theta)p(\Theta|\alpha) && (22) \\
&\propto \prod_{k=1}^{K} \Theta_k^{N_k} \Theta_k^{\alpha_k - 1} && (23) \\
&\propto \prod_{k=1}^{K} \Theta_k^{N_k + \alpha_k - 1} && (24) \\
&\propto Dir(\Theta|(\alpha_1 + N_1, \alpha_2 + N_2, \dots, \alpha_K + N_K) && (25)
\end{aligned}$$

So because Dir is normalized we conclude that

$$p(\Theta|D) = Dir(\Theta|(\alpha_1 + N_1, \alpha_2 + N_2, \dots, \alpha_K + N_K)$$

In order to derive the MAP solution for the Θ via $d/d\Theta \; p(\Theta|\mathcal{D}) \stackrel{!}{=} 0$, we have to use a Lagrange-multiplier with the constraint $\sum_k \Theta_k = 1$. The same is neccessary for the MLE solution for the Θ via $d/d\Theta \; p(\mathcal{D}|\Theta) \stackrel{!}{=} 0$. For the MLE case we get

$$\frac{\partial}{\partial \lambda}(p(\mathcal{D}|\Theta) + \lambda(1 - \sum_k \Theta_k)) \stackrel{!}{=} 0$$

and

$$\frac{\partial}{\partial \Theta_i}(p(\mathcal{D}|\Theta) + \lambda(1 - \sum_k \Theta_k)) \stackrel{!}{=} 0$$

for $i \in \{1, \ldots, K\}$.

For computing the MAP case we have to replace the likelihood $p(\mathcal{D}|\Theta)$ with the posterior $p(\Theta|\mathcal{D})$ in the upper formulae.

The results are

$$\Theta_k^{MAP} = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}$$

and

$$\Theta_k^{MLE} = \frac{N_k}{N}$$

The posterior predictive distribution for a new toss F of the K-sided dice for MLE/MAP is

$$p(F|\Theta_{MLE/MAP}) = \prod_{k=1}^{K} {\Theta_k}_{MLE/MAP}{}^{\mathbb{1}(F=k)}$$

or equivalently

$$p(F = k|\Theta_{MLE/MAP}) = {\Theta_k}_{MLE/MAP}$$

For the full Bayesian approach we have for the posterior predictive distribution:

$$\begin{aligned}
p(F = k|\mathcal{D}) &= \int_{S_K} p(F = k|\Theta)p(\Theta|D)d\Theta && (26)\\
&= \int_0^1 p(F = k|\Theta)p(\Theta|D)d\Theta_k && (27)\\
&= \int_0^1 \Theta_k Dir(\Theta|(\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)d\Theta_k && (28)\\
&= \frac{N_k + \alpha_k}{N + \alpha_0} && (29)
\end{aligned}$$

where the second line follows from "integrating out" the other $\Theta_j, j \neq k$. The K-dimensional integral of the first line must be executed over the simplex S_K.

Problem 3: Classification with a Naive Bayes classifier

In the above examples (coin, dice) we had categorial data, meaning that there was no "distinction" between a "pattern" F (H or T) and its "class" Y (heads or tail). Formally we may write $p(F = k|Y = j, \mathcal{D}) = \delta_{kj} p(F = k|\mathcal{D})$ (In other words, the probability that a coins shows "heads" but the class is "tail" is zero). In more realistic machine learning, a pattern X has a non-trivial probability of belonging to several classes so we need to distinguish pattern x and class y. Furthermore, we will usually not have one-dimensional patterns but D-dimensional patterns.

Assume that patterns e.g. are word-counts for a fixed vocabulary of size D. Thus a text-document corresponds to a pattern vector $x \in \mathbb{N}^D$. Naive Bayes models generally assume, that the features $x_j \in \mathbb{N}, j \in 1, \dots, D$ are conditionally independent given the class, so that we have for the (class-)likelihood (also called "'class-conditional density"')

$$p(x|y = c, \Theta) = \prod_{j=1}^{D} p(x_j|y = c, \Theta)$$

Let Θ denote the set of all the parameters: $\Theta = \{\pi, \theta\} = \{\pi_c, \theta_{jc}|c \in \{1, \dots, C\}, j \in \{1, \dots, D\}\}$.

So for each class, we have a class-specific set of parameters $\Theta_c = \{\theta_{jc}, \pi_c|j \in \{1, \dots, D\}\}$.

For text-classification, θ_{jc} is the probability of generating a a word j in a document in in that class c ($\leftrightarrow$ "'generative model"') assuming that $p(x|y = c, \theta) = Mu(x|N, D, \theta) = X!/(x_1!x_2! \dots x_D!) \prod_{j=1}^{D} \Theta_j^{x_j}$ where x_j is the count of word j in the document and X is the number of words in the document. As training data $\mathcal{D}$, we have a number of known vectors $x^{(i)}$ and their known class labels $y^{(i)}$: $\mathcal{D} = ((x^{(1)}, y^{(1)}), (x^{(1)}, y^{(2)}), \dots, (x^{(N)}, y^{(N)}))$.

$p(y^{(i)} = c|\pi)) = \pi_c$ denotes respective element of the vector π of prior probabilities of the classes .

- Derive the general expression for the joint probability $p(x^{(i)}, y^{(i)}|\Theta)$ of a single pattern-vector and its class-label!
- Derive a general expression for the log-likelihood $\log p(\mathcal{D}|\Theta)$!
- Derive a general expressions for the predictive posterior distribution for MLE, MAP, and the full Bayesian approach!

Solution to problem 3

Recommended reading / reference: Murphy 3.4

The joint probability of a single pattern (word-vector) and its class-label is

$$\begin{aligned}
p(x,y|\Theta) &= p(x|y,\Theta)p(y|\Theta) && (30)\\
&= p(x|y,\theta,\pi)p(y|\theta,\pi) && (31)\\
&= p(x|y,\theta)p(y|\pi) && (32)\\
&= \prod_{j=1}^{D} p(x_j|y,\theta)p(y|\pi) && (33)\\
&= \prod_{c=1}^{C}\prod_{j=1}^{D} p(x_j|\theta_{jc})^{\mathbb{1}(y=c)} \prod_{c'=1}^{C} \pi_{c'}^{\mathbb{1}(y=c')} && (34)\\
& && (35)
\end{aligned}$$

where: third line: (class-)likelihood is conditionally independent of π and (class-)prior is conditionally independent of θ_{jc}; second to last line: Naive assumption; last line: we assume a categorial distribution for the class priors (rolling a C-sided dice to determine the class).

So because of iid training examples, for N many training examples (documents) the likelihood (likelihood in terms of the parameters Θ to be learned) is

$$\begin{aligned}
p(\mathcal{D}|\Theta) &= \prod_{i=1}^{N} p(x^{(i)},y^{(i)}|\theta,\pi) && (36)\\
&= \prod_{i=1}^{N} p(x^{(i)}|y^{(i)},\theta)p(y^{(i)}|\pi) && (37)\\
&= \prod_{i=1}^{N}\prod_{j=1}^{D} p(x_j^{(i)}|y^{(i)},\theta)p(y^{(i)}|\pi) && (38)\\
&= \prod_{i=1}^{N}\prod_{c=1}^{C}\prod_{j=1}^{D} p(x_j^{(i)}|\theta_{jc})^{\mathbb{1}(y^{(i)}=c)} \prod_{c'=1}^{C} \pi_{c'}^{\mathbb{1}(y^{(i)}=c')} && (39)\\
& && (40)
\end{aligned}$$

Then the log likelihood is

$$logp(\mathcal{D}|\Theta) = \sum_{c=1}^{C}\sum_{j=1}^{D}\sum_{\{i|y^{(i)}=c\}} log\; p(x_j^{(i)}|\theta_{jc}) + \sum_{c'=1}^{C} N_{c'} log\; \pi_{c'}$$

from which we can derive the MLE estimation Θ_{MLE} for the parameters via $\partial/\partial\Theta_{jc}\; log\; p(\mathcal{D}|\theta_{jc}) \overset{!}{=} 0$ and $\partial/\partial\pi_c\; log\; p(\mathcal{D}|\Theta) \overset{!}{=} 0$

We can also derive a MAP estimate by incorporating suitable conjugate priors $p(\Theta|\alpha,\beta) = p(\theta|\beta)p(\pi|\alpha)$:

$$p(\Theta|\mathcal{D}) \propto p(\mathcal{D}|\Theta)p(\Theta|\alpha,\beta)$$

(as always: posterior(Θ) $\propto$ likelihood(Θ) * prior(Θ))

and taking the logarithm and computing the argmax.

For computing a plugin predictive posterior distribution for the class y of a new (previously unseen) pattern vector x using $\Theta_{MLE/MAP}$ we have

$$p(y = c|x, \Theta_{MAP/MLE}) \propto p(x|y = c, \Theta_{MAP/MLE}) * p(y|\Theta_{MAP/MLE})$$

(class-posterior $\propto$ class-likelihood * class-prior)

(all using the ()plugin) MAP/MLE estimate for Θ)

The likelihood (of the class label) using the MAP/MLE estimate for Θ can also be called generative class conditional density for x.)

The full Bayesian approach for the predictive posterior distribution for the class y of a new (previously unseen) pattern vector x (as always) integrates over the possible values of the Θ:

$$\begin{aligned} p(y = c|x, \mathcal{D}) &\propto \int p(x|y = c, \Theta)p(y|\Theta)p(\Theta|\mathcal{D})d\Theta && (41) \\ &\propto \int p(x|y = c, \theta, \pi)p(y|\theta, \pi)p(\theta, \pi|\mathcal{D})d\theta d\pi && (42) \\ &\propto \int p(x|y = c, \theta, \pi)p(y|\theta, \pi)p(\theta|\mathcal{D})p(\pi|\mathcal{D})d\theta d\pi && (43) \\ &\propto \int p(x|y = c, \theta)p(\theta|\mathcal{D})d\theta \int p(y|\pi)p(\pi|\mathcal{D})d\pi && (44) \end{aligned}$$

The third line corresponds to the assumption of factorized priors.

(As always) we use the Θ-posterior $p(\Theta|\mathcal{D}) = p(\mathcal{D}|\Theta)p(\Theta|\alpha, \beta)$ so very elaborately we have

$$p(y = c|x, \mathcal{D}) \propto \int p(x|y = c, \Theta)p(y|\Theta)p(\mathcal{D}|\Theta)p(\Theta|\alpha, \beta)d\Theta$$

or in words: Full Bayesian predictive posterior distribution for the class is integral over

class-likelihood * class-prior * Θ-likelihood * Θ-prior.

Reasonable choices for the involved distributions are e.g.:

- Multinomial distribution for the (class-)likelihoods / generative class conditional density: (Rolling a class-specific word-dice N_i times to create a document $x^{(i)}$)

  $$p(x^{(i)}|y^{(i)} = c, \theta) = Mu(x_1^{(i)}, \dots, x_D^{(i)}|\theta_{1c}, \dots, \theta_{Dc}) = \frac{N^{(i)}}{\prod_{j=1}^{D} x_j^{(i)}} \prod_{j=1}^{D} \theta_{jc}^{x_j^{(i)}}$$

  (Here we assume for simplicity that N_i is independent of the class. Furthermore (see Murphy p.88 remark 3 (at the end of the page)) the $x_j^{(i)}$ are not really independent ($\leftrightarrow$"Naive Bayes") because they must obey $\sum_j x_j^{(i)} \stackrel{!}{=} N_i$.)

- Categorial distribution (C-sided dice) for the class-priors (as before):

$$p(y|\pi) = \prod_{c=1}^{C} \pi_c^{(y=c)}$$

- Dirichlet distribution for the prior for θ:

$$p(\theta|\beta) = Dir(\theta|\beta)$$

- and Dirichlet-distribution for the prior for π:

$$p(\pi|\alpha) = Dir(\pi|\alpha) \tag{45}$$

A more simple choice is discussed in Murphy 3.5.2