CIS6930/4930 – Probability for Computer Systems and Machine Learning

Classification - Discrete Case

Feb. 22, 2022

Prof. Ye Xia

Problem Setup

- $X \in \{1, \dots, K\}^D$: random input vector, representing D features; each feature takes K possible values.
- $Y \in \{1, \dots, C\}$: random output scalar, representing the class.
- $P_{X,Y}(x, y)$: (joint) probability mass function, which we don't really know; short hand $P(x, y)$.
- We wish: Given an input X, we classify it into a class $\hat{Y}$ through some function h, i.e., $\hat{Y} = h(X)$.
- Example: Consider classification of a collection of documents. First, create a list of unique words that show up in the sample documents, say D words.

  A document i can be represented by a D-dimensional binary vector, say x_i, where $x_{ij} = 1$ if the jth word in the list shows up in the document, and $x_{ij} = 0$ otherwise. Example:

$$x_i = (1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1).$$

Here, there are D features, corresponding to the words. Each feature can take two possible values, 0 or 1. (Therefore, $X \in \{0, 1\}^D$)

- Let $\{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ be a random sample. The (X_i, Y_i)'s are IID, each having the same distribution as (X, Y).
- We observe a realization of the random sample, i.e., the training data: $(x_1, y_1), \ldots, (x_N, y_N)$. Each x_i is a vector corresponding to a document i; y_i is its class.

Model with Non-Random Parameters

- The strategy is to assume $P_{X,Y}$ comes from a know distribution but unknown parameters, $\pi = (\pi_c)$ and $\theta = (\theta_{jc})$, which are vectors (or matrix).
- Let $t = (t_j)$ be a feature vector, and c be a scalar for class. We have

$$P(X = t, Y = c|\pi, \theta) = P(X = t|Y = c, \pi, \theta)P(Y = c|\pi, \theta).$$

- We assume Y depends only on π:

$$P(Y = c|\pi, \theta) = P(Y = c|\pi) = \pi_c.$$

- We also assume, conditional on $Y = c$, X depends only on θ according to

$$P(X = t|Y = c, \pi, \theta) = \prod_{j=1}^{D} P(X_j = t_j|Y = c, \theta_{jc}).$$

The above says: Conditional on $Y = c$, (i) the different features are independent, and (ii) each feature j depends only on the parameter θ_{jc}. The resulting model is known as a **naive Bayes classifier**.

- For binary features, $t_j \in \{0, 1\}$ for each j. We further assume Bernoulli distributions. For each j, conditional on $Y = c$, $X_j \sim \text{Bernoulli}(\theta_{jc})$.

$$f(t_j|\theta_{jc}) \triangleq P(X_j = t_j|Y = c, \theta_{jc}) = \begin{cases} \theta_{jc}, & t_j = 1 \\ 1 - \theta_{jc}, & t_j = 0. \end{cases} \tag{1}$$

Sometimes, we use the following compact notation.

$$f(t_j|\theta_{jc}) = \theta_{jc}^{t_j}(1 - \theta_{jc})^{1-t_j}.$$

- The probability of observing a training data point (x_i, y_i) is

$$
\begin{aligned}
P(X = x_i, Y = y_i|\pi, \theta) &= P(X = x_i|Y = y_i, \pi, \theta) P(Y = y_i|\pi, \theta) \\
&= \pi_{y_i} \prod_{j=1}^{D} P(X_j = x_{ij}|Y = y_i, \theta_{jy_i}) \\
&= \pi_{y_i} \prod_{j=1}^{D} f(x_{ij}|\theta_{jy_i}) \qquad (2) \\
&= \left(\prod_{c=1}^{C} \pi_c^{\mathbf{1}(c=y_i)} \right) \left(\prod_{j=1}^{D} \prod_{c=1}^{C} f(x_{ij}|\theta_{jc})^{\mathbf{1}(c=y_i)} \right).
\end{aligned}
$$

The last step is an often used trick.

Likelihood Function

- Let the entire training data be denoted by $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$.
- The **likelihood function** is

$$P(\mathcal{D}|\pi, \theta) = \prod_{i=1}^{N} \left(\left(\prod_{c=1}^{C} \pi_c^{\mathbf{1}(c=y_i)} \right) \left(\prod_{j=1}^{D} \prod_{c=1}^{C} f(x_{ij}|\theta_{jc})^{\mathbf{1}(c=y_i)} \right) \right).$$

The above can be written as

$$\begin{aligned} P(\mathcal{D}|\pi, \theta) &= \left(\prod_{i=1}^{N} \prod_{c=1}^{C} \pi_c^{\mathbf{1}(c=y_i)} \right) \left(\prod_{i=1}^{N} \prod_{j=1}^{D} \prod_{c=1}^{C} f(x_{ij}|\theta_{jc})^{\mathbf{1}(c=y_i)} \right) \\ &= \left(\prod_{c=1}^{C} \prod_{i=1}^{N} \pi_c^{\mathbf{1}(c=y_i)} \right) \left(\prod_{j=1}^{D} \prod_{c=1}^{C} \prod_{i=1}^{N} f(x_{ij}|\theta_{jc})^{\mathbf{1}(c=y_i)} \right). \end{aligned}$$

- Note that

$$\prod_{i=1}^{N} \pi_c^{\mathbf{1}(c=y_i)} = \pi_c^{\sum_{i=1}^{N} \mathbf{1}(c=y_i)} = \pi_c^{N_c},$$

where, for each c, $N_c \triangleq \sum_{i=1}^{N} \mathbf{1}(c = y_i)$ is the number of data points that belong to class c.

- Next, consider

$$\prod_{i=1}^{N} f(x_{ij}|\theta_{jc})^{\mathbf{1}(c=y_i)}.$$

Here, j and c are fixed. The above is a product of N factors. For i such that $x_{ij} = 1$ and $c = y_i$, the factor is θ_{jc}; for i such that $x_{ij} = 0$ and $c = y_i$, the factor is $(1 - \theta_{jc})$; for i such that $c \neq y_i$, the factor is 1. Thus,

$$\prod_{i=1}^{N} f(x_{ij}|\theta_{jc})^{\mathbf{1}(c=y_i)} = \theta_{jc}^{\sum_{i=1}^{N} \mathbf{1}(c=y_i, x_{ij}=1)} (1-\theta_{jc})^{\sum_{i=1}^{N} \mathbf{1}(c=y_i, x_{ij}=0)}.$$

- For each j and c, let $N_{jc} \triangleq \sum_{i=1}^{N} \mathbf{1}(x_{ij} = 1, c = y_i)$.

  N_{jc} is the number of data points that belong to class c and with the jth feature turned on.

- Then, $N_c - N_{jc}$ is the number of data points that belong to class c and with the jth feature turned off. That is,
  $N_c - N_{jc} = \sum_{i=1}^{N} \mathbf{1}(x_{ij} = 0, c = y_i)$.

- We can write

$$\prod_{i=1}^{N} f(x_{ij}|\theta_{jc})^{\mathbf{1}(c=y_i)} = \theta_{jc}^{N_{jc}}(1 - \theta_{jc})^{N_c - N_{jc}}.$$

- The likelihood function can be written as

$$P(\mathcal{D}|\pi, \theta) = \left(\prod_{c=1}^{C} \pi_c^{N_c}\right)\left(\prod_{j=1}^{D}\prod_{c=1}^{C} \theta_{jc}^{N_{jc}}(1 - \theta_{jc})^{N_c - N_{jc}}\right). \quad (3)$$

Log-Likelihood Function

- The **log-likelihood function** is

$$\log P(\mathcal{D}|\pi, \theta)$$

$$= \sum_{c=1}^{C} N_c \log \pi_c + \sum_{c=1}^{C} \sum_{j=1}^{D} \left(N_{jc} \log \theta_{jc} + (N_c - N_{jc}) \log(1 - \theta_{jc}) \right). \tag{4}$$

Maximum Likelihood Estimate (MLE)

- Choose π and θ that maximizes the likelihood $P(\mathcal{D}|\pi, \theta)$. The maximizing (π, θ), denoted by $(\hat{\pi}, \hat{\theta})$, is an MLE.
- Maximizing $P(\mathcal{D}|\pi, \theta)$ is the same as maximizing $\log P(\mathcal{D}|\pi, \theta)$.
- For (4), due to the separation of the variables, we ends up with two sets of maximization sub-problems.
- The first is

$$\max_{\pi} \sum_{c=1}^{C} N_c \log \pi_c$$

$$\text{subject to } \sum_{c=1}^{C} \pi_c = 1.$$

This is an easy problem. The maximizer is, for each c,

$$\hat{\pi}_c = \frac{N_c}{N}.$$

- $\hat{\pi}_c$ is simply the empirical average.
- Note that if $N_c = 0$ (possibly due to insufficient data in high-dimensional cases), $\hat{\pi}_c = 0$. Class c is ruled out. This is known as the **zero count problem**, a form of overfitting.

- The second set of sub-problems is, for each j and c,

$$\max_{\theta_{jc} \in [0,1]} N_{jc} \log \theta_{jc} + (N_c - N_{jc}) \log(1 - \theta_{jc}).$$

- First, assume $0 < N_{jc} < N_c$. Taking derivative with respect to θ_{jc}

and setting it to zero, we get

$$\frac{N_{jc}}{\theta_{jc}} - \frac{N_c - N_{jc}}{1 - \theta_{jc}} = 0.$$

This yields the maximizer

$$\hat{\theta}_{jc} = \frac{N_{jc}}{N_c}.$$

- $\hat{\theta}_{jc}$ is again an empirical average.
- For cases of $N_{jc} = 0$ or $N_{jc} = N_c$, the above solution is still correct.
- The MLE is random because it depends on the random sample $\{(X_1, Y_1), \ldots, (X_N, Y_N)\}$.

How to Classify New Document

- With $\hat{\pi}, \hat{\theta}$, we now think the distribution of (X, Y) is $P(t, c|\hat{\pi}, \hat{\theta})$ with the form given in (2), i.e.,

$$P(X = t, Y = c|\hat{\pi}, \hat{\theta}) = \hat{\pi}_c \prod_{j=1}^{D} f(t_j|\hat{\theta}_{jc}),$$

  where f is the Bernoulli pmf given in (1).

  Keep in mind that this is still not the true distribution of (X, Y).

- The marginal probability mass for X is

$$P(X = t|\hat{\pi}, \hat{\theta}) = \sum_{c=1}^{C} \hat{\pi}_c \prod_{j=1}^{D} f(t_j|\hat{\theta}_{jc}).$$

- For a given document with the feature vector t, the class that it belongs to is random, according to the conditional probability

$$P(Y = c|X = t, \hat{\pi}, \hat{\theta}) = \frac{P(X = t, Y = c|\hat{\pi}, \hat{\theta})}{P(X = t|\hat{\pi}, \hat{\theta})}, \quad c \in \{1, \ldots, C\}.$$

- If we insist on putting t into one class, we can use the mode as its class. Let
$$\hat{c} \in \operatorname*{argmax}_c P(Y = c|X = t, \hat{\pi}, \hat{\theta}).$$
Then, our classification function is $h(t) = \hat{c}$.

- Since $P(X = t|\hat{\pi}, \hat{\theta})$ does not depend on c, $\hat{c}$ is also the mode of $P(X = t, Y = c|\hat{\pi}, \hat{\theta})$. We only need to find the mode for $P(X = t, Y = c|\hat{\pi}, \hat{\theta})$ but do not need to compute the marginal probability $P(X = t|\hat{\pi}, \hat{\theta})$.

- For numerical calculation, it is easier to compute $\log P(X = t, Y = c|\hat{\pi}, \hat{\theta})$ for different c and find the maximizing $\hat{c}$.

Problem of MLE - Overfitting

- Consider email documents. Suppose the jth feature corresponding to the word ‘subject’, and suppose $\hat{\theta}_{jc} = 1$. This happens if all the training documents contain the word ‘subject’ (since $\hat{\theta}_{jc} = \frac{N_{jc}}{N_c}$)
- In the classification stage, consider an input document without the word ‘subject’. The input vector t has $t_j = 0$. Then, $f(t_j|\hat{\theta}_{jc}) = 1 - \hat{\theta}_{jc} = 0$ for all c. We have, for every c,

$$P(X = t, Y = c|\hat{\pi}, \hat{\theta}) = \hat{\pi}_c \prod_{j=1}^{D} f(t_j|\hat{\theta}_{jc}) = 0.$$

- Classification will fail.
- A Bayesian approach will solve the problem.

Bayesian Approach

- In the Bayesian approach, we don't think π and θ as unknown parameters. We think about them as random variables Π and Θ, with a prior distribution.
- The inference will be based on the posterior probability.
- **Assumption 0:** For simplicity, we assume a factored prior:

$$P_{\Pi,\Theta}(\pi,\theta) = P_{\Pi}(\pi) \prod_{j=1}^{D} \prod_{c=1}^{C} P_{\Theta_{jc}}(\theta_{jc}).$$

  When there is no confusion, we will omit the subscripts.

  Note that $\{\Theta_{jc}\}_{jc}$ are independent of each other, and independent of all π_c; $\{\Pi_c\}_c$ may be dependent on each other.

- **Assumption 1:** As before, we still assume

$$P(Y = c|\pi, \theta) = P(Y = c|\pi) = \pi_c. \tag{5}$$

But, the meaning is different since Π and θ are random variables. The first equality says that conditional on $\Pi = \pi$, Y is independent of Θ.

- **Assumption 2:** As before, we still assume,

$$P(X = t|Y = c, \pi, \theta) = \prod_{j=1}^{D} P(X_j = t_j|Y = c, \theta_{jc}), \tag{6}$$

where each $P(X_j = t_j|Y = c, \theta_{jc})$ is a function of t_j and θ_{jc} only, in particular, the Bernoulli pmf $f(t_j|\theta_{jc})$ in (1).

This implies X is independent of Π given Y and Θ (prove this):

$$P(X = t|Y = c, \pi, \theta) = P(X = t|Y = c, \theta). \tag{7}$$

Proof:

$$
\begin{aligned}
P(X=t|Y=c,\theta) &= \int P(X=t|Y=c,\pi,\theta)P(\pi|Y=c,\theta)d\pi \\
&= \int \prod_{j=1}^{D} P(X_j=t_j|Y=c,\theta_{jc})P(\pi|Y=c,\theta)d\pi \\
&= \prod_{j=1}^{D} P(X_j=t_j|Y=c,\theta_{jc}) \int P(\pi|Y=c,\theta)d\pi \\
&= \prod_{j=1}^{D} P(X_j=t_j|Y=c,\theta_{jc}) \\
&= P(X=t|Y=c,\pi,\theta).
\end{aligned}
$$

The integrals in the above are over $\pi_1, \ldots, \pi_C$ in the region where $\pi_c \geq 0$ for each c and $\sum_{c=1}^{C} \pi_c = 1$.

- **Assumption 3:** For each π and θ, conditional on $\{\Pi = \pi, \Theta = \theta\}$,

$(X_i, Y_i)_{i=1}^N$ in the sample are IID.

- Under the above assumptions, the likelihood function is as in (3):

$$P(\mathcal{D}|\pi, \theta) = \left(\prod_{c=1}^{C} \pi_c^{N_c} \right) \left(\prod_{j=1}^{D} \prod_{c=1}^{C} \theta_{jc}^{N_{jc}} (1 - \theta_{jc})^{N_c - N_{jc}} \right).$$

Recall $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$.

MAP - Maximum A Posteriori

- One of our goals is to find (π, θ) that maximizes the posterior probability $P(\pi, \theta|\mathcal{D})$. A maximizer $(\hat{\pi}, \hat{\theta})$ is known as the **MAP estimate**.
- Note that

$$P(\pi, \theta|\mathcal{D}) = \frac{P(\mathcal{D}|\pi, \theta)P(\pi, \theta)}{P(\mathcal{D})}.$$

- $P(\mathcal{D})$ doesn't contain π or θ, and thus it plays no role in the maximization. We can ignore $P(\mathcal{D})$ and write:

$$P(\pi, \theta|\mathcal{D}) \propto P(\mathcal{D}|\pi, \theta)P(\pi, \theta).$$

- Then,

$$P(\pi, \theta|\mathcal{D}) \propto$$

$$\left(P(\pi)\prod_{j=1}^{D}\prod_{c=1}^{C}P(\theta_{jc})\right)\left(\prod_{c=1}^{C}\pi_c^{N_c}\right)\left(\prod_{j=1}^{D}\prod_{c=1}^{C}\theta_{jc}^{N_{jc}}(1-\theta_{jc})^{N_c-N_{jc}}\right).$$

- After re-arranging,

$$P(\pi, \theta|\mathcal{D}) \propto \left(P(\pi) \prod_{c=1}^{C} \pi_c^{N_c} \right) \left(\prod_{j=1}^{D} \prod_{c=1}^{C} \left(P(\theta_{jc}) \theta_{jc}^{N_{jc}} (1-\theta_{jc})^{N_c - N_{jc}} \right) \right).$$

We see that the π-part and θ-part are separated. Also, each θ_{jc} is separated from each other as well.

- **Main result:** We need to solve a set of smaller maximization sub-problems, one for π and one for each θ_{jc}.

Sub-Problem 1:

$$\max_{\pi} P(\pi) \prod_{c=1}^{C} \pi_c^{N_c}$$

$$\text{s.t. } \pi \geq 0, \sum_{c=1}^{C} \pi_c = 1$$

Sub-Problems 2: For each j, c,

$$\max_{0 \leq \theta_{jc} \leq 1} P(\theta_{jc}) \theta_{jc}^{N_{jc}} (1-\theta_{jc})^{N_c - N_{jc}}$$

Derivation based on Factored Model

- This separation result can be derived from the general results about factored models (see later). Since the likelihood function and the prior can both be factorized into the π part and θ part, the general results lead to factorized posterior:

$$P(\pi, \theta|\mathcal{D}) = P(\pi|\mathcal{D})P(\theta|\mathcal{D}). \tag{8}$$

- The model is also factorized with respect to θ_{jc}. Thus, we have

$$P(\pi, \theta|\mathcal{D}) = P(\pi|\mathcal{D}) \prod_{j=1}^{D} \prod_{c=1}^{C} P(\theta_{jc}|\mathcal{D}).$$

Now,

$$\max_{\pi,\theta} P(\pi, \theta|\mathcal{D}) = \max_{\pi} P(\pi|\mathcal{D}) \prod_{j=1}^{D} \prod_{c=1}^{C} \max_{\theta_{jc}} P(\theta_{jc}|\mathcal{D}).$$

- Also due to the factored model (see later) and the particular expression of the likelihood function, we have

$$P(\pi|\mathcal{D}) \propto P(\pi) \prod_{c=1}^{C} \pi_c^{N_c}$$

$$P(\theta_{jc}|\mathcal{D}) \propto P(\theta_{jc}) \theta_{jc}^{N_{jc}} (1 - \theta_{jc})^{N_c - N_{jc}}, \ \ \forall j, c$$

- We see that $\max P(\pi|\mathcal{D})$ leads to sub-problem 1 earlier; $\max P(\theta_{jc}|\mathcal{D})$ leads to sub-problems 2.

What Prior Distributions?

- There is much flexibility when the training data points are sufficient.
- **Here is a general discussion with new notations.**
- To simplify the notation, suppose the parameters are all collected into a single random vector Θ (no Π anymore). Suppose the random sample is $X_1, \ldots, X_N$ (no Y anymore).
- Suppose the training data points are $\mathcal{D} = (x_1, x_2, \ldots, x_N)$, where each x_i is a vector in general.
- We have

$$P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta) = \prod_{i=1}^{N} P_{X|\Theta}(x_i|\theta)P(\theta).$$

Then,

$$\log P(\theta|\mathcal{D}) \propto \sum_{i=1}^{N} \log P_{X|\Theta}(x_i|\theta) + \log P(\theta).$$

The MAP estimate is

$$\operatorname*{argmax}_{\theta} \left(\sum_{i=1}^{N} \log P_{X|\Theta}(x_i|\theta) + \log P(\theta) \right).$$

- The term $\sum_{i=1}^{N} \log P_{X|\Theta}(x_i|\theta)$ is roughly linear in N. As N is sufficiently large, it overwhelms the term $\log P(\theta)$. If so, the MAP estimate converges to the MLE.
- When there is enough data, we say the **data overwhelms the prior**.

Conjugate Prior and Beta-Binomial Model

- This part is a general discussion.
- For easy calculation, one often chooses a conjugate prior.
- When the prior is such that the prior and the posterior have the same form, we say the prior is a **conjugate prior** for the corresponding likelihood function.
- Suppose the random sample $X_1, \ldots, X_N$ is a sequence of IID Bernoulli random variables with parameter θ (a scalar). Let $\mathcal{D} = (x_1, \ldots, x_N)$ be a realization (i.e., data).
- Then, the likelihood function is

$$P(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0},$$

where $N_1 = \sum_{i=1}^{N} \mathbf{1}(x_1 = 1)$ and $N_0 = N - N_1$.

- The log-likelihood function is

$$\log P(\mathcal{D}|\theta) = N_1 \log \theta + N_0 \log(1-\theta).$$

The MLE is $\hat{\theta} \in \mathrm{argmax}_{\theta \in [0,1]} \log P(\mathcal{D}|\theta)$. This gives

$$\hat{\theta}_{MLE} = \frac{N_1}{N}. \tag{9}$$

- A conjugate prior has the following form for its pdf:

$$P(\theta) \propto \theta^{\gamma_1}(1-\theta)^{\gamma_2}, \;\; \theta \in [0,1]. \tag{10}$$

- With that, the posterior is:

$$P(\theta|\mathcal{D}) \propto \theta^{\gamma_1}(1-\theta)^{\gamma_2}\theta^{N_1}(1-\theta)^{N_0} = \theta^{N_1+\gamma_1}(1-\theta)^{N_0+\gamma_2}. \tag{11}$$

Beta Distribution

- A beta distribution with parameters a and b is denoted by $\text{Beta}(a, b)$. The pdf is

$$\text{Beta}(\theta; a, b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1},\ \theta \in [0,1],$$

where $a, b > 0$ and $B(a, b)$ is the **beta function** defined by

$$B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta.$$

- $B(a, b)$ is related to the gamma function by

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Recall a gamma function is $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ for $t > 0$.

- For a random variable X with the distribution Beta(a, b):

$$E[X] = \frac{a}{a+b}, \quad \text{var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

- **Mode:** When $a > 1$ and $b > 1$, it has a single mode at

$$\frac{a-1}{a+b-2} \qquad \text{(most important case)}$$

When $a < 1$ and $b < 1$: two modes 0 and 1

When $a \leq 1$ and $b > 1$: 0

When $a > 1$ and $b \leq 1$: 1

When $a = b = 1$: uniform distribution on $[0, 1]$.

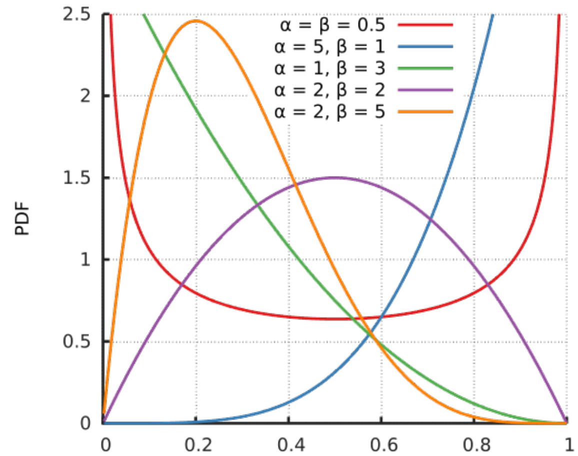

Figure 1: Examples of beta pdf (from Wikipedia)

- As we increase a and b to infinity, we see that the variance decreases to zero, and therefore, the pdf is more and more peaked.

Where Is the Binomial Part?

- We see that the likelihood function $P(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$ depends on the data $\mathcal{D}$ through N_1 only ($N_0 = N - N_1$ is a function of N_1).
- For the purpose of estimating θ, we only need to collect N_1 from the data.
- We will learn later that N_1 is a sufficient statistic (in fact, a realization of it) with respect to the estimation of θ.
- For our random sample $X_1, \ldots, X_N$, let $S(X_1, \ldots, X_N)$ be a statistic (i.e., a function of the random sample). Let $\mathcal{D} = (x_1, \ldots, x_N)$ be a realization of the random sample.
- According to one definition, $S(X_1, \ldots, X_N)$ is a **sufficient statistic** if $P(\theta|\mathcal{D}) = P(\theta|S(\mathcal{D}))$ for all θ and all $\mathcal{D}$.
- In our case, the sufficient statistic involved is

$S(X_1, \ldots, X_N) = \sum_{i=1}^{N} \mathbf{1}(X_i = 1)$, which has a binomial distribution:

$$P(S(X_1, \ldots, X_N) = N_1) = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N-N_1} = \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_0}.$$

- We see that, for any $\mathcal{D}$ with N_1 successes, i.e., $S(\mathcal{D}) = N_1$, we have

$$\begin{aligned} P(\theta|\mathcal{D}) &= P(\theta|S(\mathcal{D})) \\ &\propto P(S(\mathcal{D})|\theta)P(\theta) \\ &= \binom{N}{N_1} \theta^{N_1} (1-\theta)^{N_0} \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}. \end{aligned}$$

- Therefore, in the beta-binomial model, the posterior is proportional to the product of a beta prior (a pdf) and a binomial distribution (a pmf). Therefore, the posterior is another beta distribution.
- We can also explain the model name without mentioning the sufficient statistic. For any $\mathcal{D}$ with N_1 successes, the likelihood function is:

$$P(\mathcal{D}|\theta) = \theta^{N_1} (1-\theta)^{N_0}.$$

- Then,

$$
\begin{aligned}
P(\theta|\mathcal{D}) &\propto P(\mathcal{D}|\theta)P(\theta) \\
&= \theta^{N_1}(1-\theta)^{N_0}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_0}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}.
\end{aligned}
$$

MAP Estimate of θ

- The conjugate prior $P(\theta)$ in (10) is $\mathrm{Beta}(a, b)$ with $a = \gamma_1 + 1$ and $b = \gamma_2 + 1$.
- A common practice is to choose $\gamma_1 = \gamma_2 = 1$ so that the prior is $\mathrm{Beta}(2, 2)$;
- With the conjugate prior, the posterior in (11) can be written as

$$P(\theta|\mathcal{D}) \propto \theta^{a-1}(1-\theta)^{b-1}\theta^{N_1}(1-\theta)^{N_0} = \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}. \tag{12}$$

  The posterior has the Beta distribution $\mathrm{Beta}(N_1 + a, N_0 + b)$.
- The MAP estimate $\hat{\theta}_{MAP}$ is the mode of the posterior $P(\theta|\mathcal{D})$:

$$\hat{\theta}_{MAP} = \frac{N_1 + a - 1}{N + a + b - 2}.$$

- Also note that the maximization problem for the MAP estimate has the same form as the one for the MLE that leads to (9). The MLE is a special case, which corresponds to the uniform prior (with $a = b = 1$).

$$\hat{\theta}_{MLE} = \frac{N_1}{N}.$$

- When the training data is such that $N_1 = 0$, we have $\hat{\theta}_{MLE} = 0$. This is the zero count problem.
- When $N_1 = N$, we have $\hat{\theta}_{MLE} = 1$. Earlier, we have identified an overfitting problem when using the MLE for classification, which is a problem of the same kind.
- Instead, suppose we use the MAP estimate and and suppose we choose the prior with $a = b = 2$. Then $\hat{\theta}_{MAP} > 0$ and we have solved the zero count problem. Furthermore, even when $N_1 = N$, we have $\hat{\theta}_{MAP} < 1$; we have solved the earlier overfitting problem.

Trading Off Prior and MLE

- Let $\alpha_0 = a + b$. One can view $\alpha_0 - 2$ as the equivalent sample size of the prior, based on the form of the posterior $P(\theta|\mathcal{D})$ in (12).
- Let the prior mean be denoted by $m_1 = \frac{a}{a+b} = \frac{a}{\alpha_0}$.
- Since the posterior is Beta$(N_1 + a, N_0 + b)$, its mean is

$$E[\theta|\mathcal{D}] = \frac{N_1 + a}{N_1 + a + N_0 + b} = \frac{N_1 + a}{N + \alpha_0}.$$

We see that the posterior mean is not the same as the posterior mode.

- Furthermore,

$$\begin{aligned} E[\theta|\mathcal{D}] =& \frac{a}{N+\alpha_0} + \frac{N_1}{N+\alpha_0} \\ =& \frac{\alpha_0}{N+\alpha_0} m_1 + \frac{N}{N+\alpha_0}\frac{N_1}{N} \\ =& \lambda m_1 + (1-\lambda)\hat{\theta}_{MLE}, \end{aligned}$$

  where $\lambda = \frac{\alpha_0}{N+\alpha_0}$ is the weight of the prior. We see that as $N \to \infty$, the posterior mean approaches the MLE.

- Similarly, one can show that the posterior mode is a convex combination of the prior mode and the MLE and that it converges to the MLE.

Dirichlet-Multinomial Model

- This is a generalization from binary to K-outcome trials.
- In this model, the prior is a Dirichlet distribution, and the likelihood function is proportional to a multinomial distribution. The posterior is another Dirichlet distribution.
- Consider a sequence of N IID random variables, $X_1, \ldots, X_N$, with

$$P(X_i = k) = \theta_k, \;\; k \in \{1, \ldots, K\},$$

where $\sum_{i=1}^{K} \theta_k = 1$. Let $\theta = (\theta_1, \ldots, \theta_K)$.

- Let the sample data points be $\mathcal{D} = \{x_1, \ldots, x_N\}$, where each $x_i \in \{1, \ldots, K\}$.
- Let $N_k = \sum_{i=1}^{N} \mathbf{1}(x_i = k)$, which is the number of times outcome k shows up.

- The likelihood function is

$$P(\mathcal{D}|\theta) = \prod_{k=1}^{K} \theta_k^{N_k}.$$

- The MLE is

$$\hat{\theta}_{MLE} \in \underset{\theta \in S_K}{\operatorname{argmax}} \log P(\mathcal{D}|\theta).$$

- The maximization is taken over the set S_K known as the K**-simplex**:

$$S_K \triangleq \{(z_1, \ldots, z_K) \in \mathbb{R}^K : \sum_{k=1}^{K} z_k = 1; z_k \geq 0, \forall k\}.$$

  Each point in S_K can be a probability assignment.

- The maximum is

$$\hat{\theta}_{MLE,k} = \frac{N_k}{N}, \quad k = 1, \ldots, K. \tag{13}$$

Dirichlet Distribution

- The **Dirichlet distribution** is a conjugate prior for the above $P(\mathcal{D}|\theta)$. It can be viewed as a generalization to the beta distribution.
- A Dirichlet distribution is a joint distribution of K random variables $Y_1, \ldots, Y_K$ with each $Y_k \geq 0$ and $\sum_{k=1}^{K} Y_k = 1$. That is, it is a distribution on the K-simplex.
- A Dirichlet distribution has K parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$, where each $\alpha_k > 0$.
- A Dirichlet distribution is denoted by $\mathrm{Dir}(\alpha)$. The pdf is

$$\mathrm{Dir}(\theta; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \;\; \theta \in S_K,$$

where $B(\alpha)$ is a generalization of the beta function to K dimension.

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)}, \qquad \alpha_0 \triangleq \sum_{k=1}^{K} \alpha_k.$$

- Suppose the random vector $Y = (Y_1, \ldots, Y_K)$ has the distribution $\mathrm{Dir}(\alpha)$. Then, for each k,

$$E[Y_k] = \frac{\alpha_k}{\alpha_0}, \qquad \mathrm{var}(Y_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}.$$

When $\alpha_k > 1$ for every k,

$$\mathrm{mode}(Y_k) = \frac{\alpha_k - 1}{\alpha_0 - K}, \;\; \forall k. \tag{14}$$

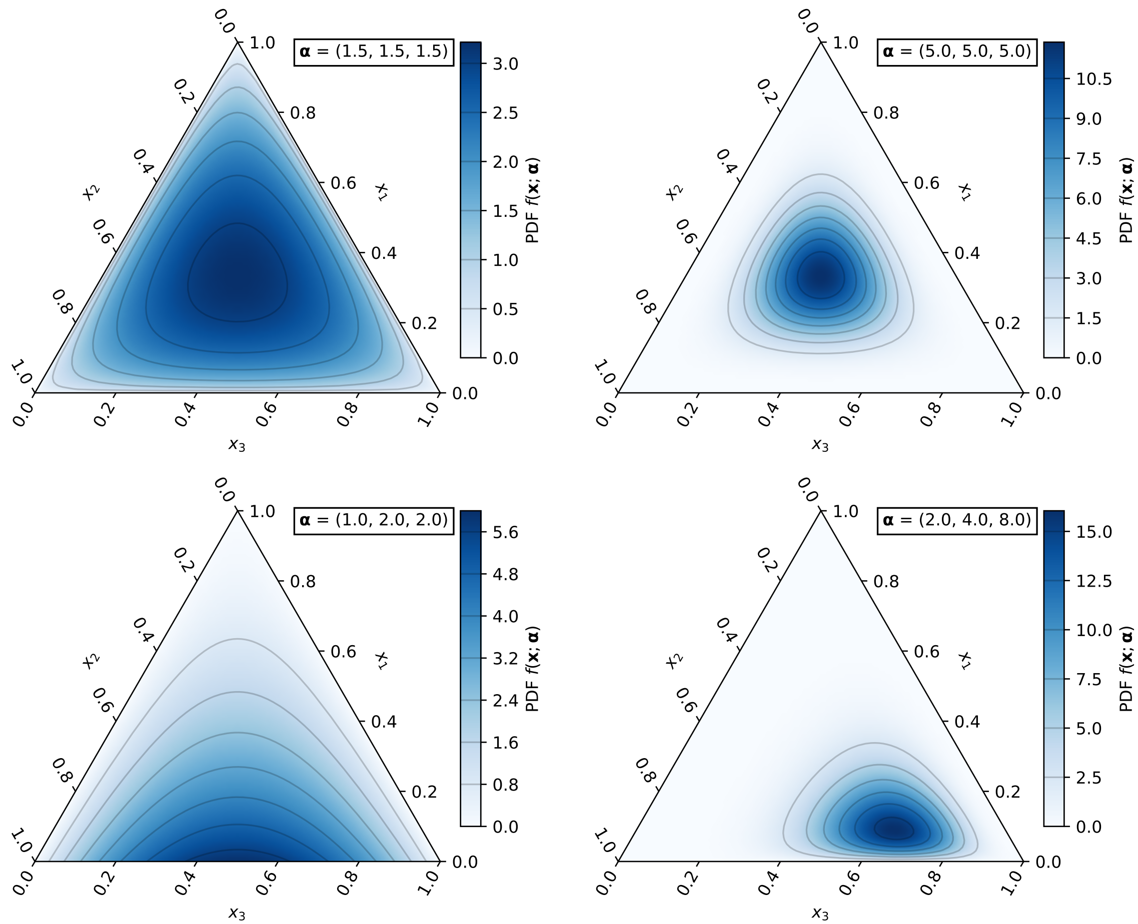

Figure 2: Examples of Dirichlet pdf (from Wikipedia)

- $\mathrm{Dir}(1,1,1)$ is the uniform distribution on the simplex.
- In practice, one often sets the Dirichlet parameters to be $\alpha_k = b/K$ for each k, where the constant $b > K$. In this case, for each k,

$$E[Y_k] = \frac{1}{K}, \qquad \mathrm{var}(Y_k) = \frac{K-1}{K^2(b+1)}.$$

  The pdf is more peaked when b is larger.
- With a Dirichlet prior $\mathrm{Dir}(\alpha)$, the posterior satisfies

$$P(\theta|\mathcal{D}) \propto P(\theta)P(\mathcal{D}|\theta) = \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{k=1}^{K} \theta_k^{N_k} = \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}.$$

  Thus, $P(\theta|\mathcal{D})$ again corresponds to a Dirichlet distribution, $\mathrm{Dir}(N_1 + \alpha_1, \ldots, N_K + \alpha_K)$.
- The MAP estimate is the mode of the posterior, which is given in

(14). Therefore, the MAP estimate is

$$\hat{\theta}_{MAP,k} = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}, \quad k = 1, \ldots, K.$$

- Also, the maximization problem has the same form as the one for the MLE that leads to (13).
- As for the multinomial part, the likelihood function is proportional to the multinomial distribution (when viewed as a function of θ).

$$P(\mathcal{D}|\theta) = \prod_{k=1}^{K} \theta_k^{N_k} \propto \frac{N!}{N_1! \ldots N_k!} \prod_{k=1}^{K} \theta_k^{N_k}.$$

The latter is the probability of the event that, after N trials, there are exactly N_k trials that have the outcome k, for every k.

- $N_1, \ldots, N_K$ are realization of a (jointly) sufficient statistic for the purpose of estimating θ. There is no need to record all the data points in $\mathcal{D}$. It is enough to collect $N_1, \ldots, N_K$ from the data points.

Back to MAP Estimate for the Classification Problem

To find the MAP estimate, we have two sets of sub-problems:

Sub-Problem 1:

$$\max_{\pi} P(\pi) \prod_{c=1}^{C} \pi_c^{N_c}$$

$$\text{s.t. } \pi \geq 0, \sum_{c=1}^{C} \pi_c = 1$$

Sub-Problems 2: For each j, c,

$$\max_{0 \leq \theta_{jc} \leq 1} P(\theta_{jc}) \theta_{jc}^{N_{jc}} (1 - \theta_{jc})^{N_c - N_{jc}}$$

- We will use the conjugate priors.
- $P(\pi)$: Dirichlet distribution $\text{Dir}(\alpha)$, e.g., $\alpha = (2, 2, \ldots, 2)$.
- Then, $P(\pi|\mathcal{D})$ is a Dirichlet distribution, $\text{Dir}(N_1 + \alpha_1, \ldots, N_C + \alpha_C)$.

  When $\alpha_c > 1$ for all c, $P(\pi|\mathcal{D})$ has a single mode, which corresponds to the MAP estimate for π.
- Let the mode be denoted by $\hat{\pi}_{MAP}$. Then, for each c,

$$\hat{\pi}_{MAP,c} = \frac{N_c + \alpha_c - 1}{N + \alpha_0 - C}, \tag{15}$$

  where $\alpha_0 = \sum_{c=1}^{C} \alpha_c$.

- $P(\theta_{jc})$ has a beta distribution $\text{Beta}(\beta_0, \beta_1)$, e.g., $\beta_0 = 2, \beta_1 = 2$.
- Then, $P(\theta_{jc}|\mathcal{D})$ is another Beta distribution, $\text{Beta}(N_{jc} + \beta_0, N_c - N_{jc} + \beta_1)$.

- When $\beta_0 > 1$ and $\beta_1 > 1$, $P(\theta_{jc}|\mathcal{D})$ has a single mode, which corresponds to the MAP estimate, to be denoted by $\hat{\theta}_{MAP,jc}$. For each j and c, it is

$$\hat{\theta}_{MAP,jc} = \frac{N_{jc} + \beta_0 - 1}{N_c + \beta_0 + \beta_1 - 2}. \tag{16}$$

Use the Model for Prediction (Classification)

- Given a new document feature vector x, our goal is to classify it.
- One possibility: We use the MAP estimate in (15) and (16), and make prediction based on

$$P(Y = c|X = x, \hat{\pi}_{MAP}, \hat{\theta}_{MAP}).$$

  Whichever c that maximizes the above probability will be the assigned class for the document. This is known as the **plug-in approximation**.

- Note that

$$
\begin{aligned}
& P(Y = c|X = x, \hat{\pi}_{MAP}, \hat{\theta}_{MAP}) \\
= & \frac{P(X = x, Y = c|\hat{\pi}_{MAP}, \hat{\theta}_{MAP})}{P(X = x|\hat{\pi}_{MAP}, \hat{\theta}_{MAP})} \\
= & \frac{P(X = x|Y = c, \hat{\pi}_{MAP}, \hat{\theta}_{MAP}) P(Y = c|\hat{\pi}_{MAP}, \hat{\theta}_{MAP})}{P(X = x|\hat{\pi}_{MAP}, \hat{\theta}_{MAP})} \\
= & \frac{\prod_{j=1}^{D} \hat{\theta}_{MAP,jc}^{x_j} (1 - \hat{\theta}_{MAP,jc})^{1-x_j} \hat{\pi}_{MAP,c}}{P(X = x|\hat{\pi}_{MAP}, \hat{\theta}_{MAP})}.
\end{aligned}
$$

To go directly from the first expression to the third, we can apply (the conditional version of) Bayes' formula.

- Since the denominator does not depend on c, we have

$$\begin{aligned} & P(Y = c | X = x, \hat{\pi}_{MAP}, \hat{\theta}_{MAP}) \\ & \propto \hat{\pi}_{MAP,c} \prod_{j=1}^{D} \hat{\theta}_{MAP,jc}^{x_j} (1 - \hat{\theta}_{MAP,jc})^{1-x_j}. \end{aligned} \tag{17}$$

- We only need to find a maximizer for the expression in (17). In fact, we will first take log and find a maximizer for log of the expression.

$$\underset{c}{\operatorname{argmax}} \left(\log \hat{\pi}_{MAP,c} + \sum_{j=1}^{D} \left(x_j \log \hat{\theta}_{MAP,jc} + (1 - x_j) \log(1 - \hat{\theta}_{MAP,jc}) \right) \right).$$

- The plug-in approach is fine. But, it does not make use all the information from the data $\mathcal{D}$. We next consider a different approach.

Classification via Posterior Predictive Distribution

- The class that a document with feature vector x belongs to is a random variable with the conditional distribution:

$$P(Y = c|X = x, \mathcal{D}) = \frac{P(Y = c|\mathcal{D})P(X = x|Y = c, \mathcal{D})}{P(X = x|\mathcal{D})}.$$

- Since $P(X = x|\mathcal{D})$ does not depend on c,

$$P(Y = c|X = x, \mathcal{D}) \propto P(Y = c|\mathcal{D})P(X = x|Y = c, \mathcal{D}). \quad (18)$$

- We will work on $P(Y = c|\mathcal{D})$ first. The following is a multiple integral with variables $\pi_1, \ldots, \pi_C$ over the simplex S_C.

$$
\begin{aligned}
P(Y = c|\mathcal{D}) &= \int P(Y = c|\pi, \mathcal{D}) P(\pi|\mathcal{D}) d\pi \\
&= \int P(Y = c|\pi) P(\pi|\mathcal{D}) d\pi \\
&= \int \pi_c P(\pi|\mathcal{D}) d\pi = E[\Pi_c|\mathcal{D}].
\end{aligned}
$$

We have used conditional independence of the sample, which leads to

$$
\text{(prove this)} \qquad P(Y = c|\pi, \mathcal{D}) = P(Y = c|\pi). \tag{19}
$$

Note that Y and the random sample that produces $\mathcal{D}$ are not independent. But, conditional on $\Pi = \pi$, they are independent. (Θ does not matter due to the factored form of the likelihood function.)

Proof of (19):

$$
\begin{aligned}
P(Y=c|\pi,\mathcal{D}) &= \frac{P(Y=c,\mathcal{D}|\pi)}{P(\mathcal{D}|\pi)} \\
&= \int \frac{P(Y=c,\mathcal{D}|\pi,\theta)P(\theta|\pi)}{P(\mathcal{D}|\pi)} d\theta \\
\text{(model Assumption 3)} &= \int \frac{P(Y=c|\pi,\theta)P(\mathcal{D}|\pi,\theta)P(\theta|\pi)}{P(\mathcal{D}|\pi)} d\theta \\
\text{(model Assumption 1)} &= \int \frac{P(Y=c|\pi)P(\mathcal{D}|\pi,\theta)P(\theta|\pi)}{P(\mathcal{D}|\pi)} d\theta \\
&= \frac{P(Y=c|\pi)}{P(\mathcal{D}|\pi)} \int P(\mathcal{D}|\pi,\theta)P(\theta|\pi)d\theta \\
&= \frac{P(Y=c|\pi)}{P(\mathcal{D}|\pi)} \int P(\mathcal{D},\theta|\pi)d\theta \\
&= \frac{P(Y=c|\pi)}{P(\mathcal{D}|\pi)} P(\mathcal{D}|\pi) = P(Y=c|\pi).
\end{aligned}
$$

- Recall that, with the conjugate prior, the posterior $P(\pi|\mathcal{D})$ is $\text{Dir}(N_1 + \alpha_1, \ldots, N_C + \alpha_C)$. For such a Dirichlet distribution, the mean is

$$\bar{\pi}_c \triangleq E[\Pi_c|\mathcal{D}] = \frac{N_c + \alpha_c}{N + \alpha_0},$$

where $\alpha_0 = \sum_{c=1}^{C} \alpha_c$.

- To summarize, we have

$$P(Y = c|\mathcal{D}) = E[\Pi_c|\mathcal{D}] = \bar{\pi}_c. \tag{20}$$

- This is understandable. To compute $P(Y = c)$, we have to do 'averaging' over the random parameter Π_c. Since we observed $\mathcal{D}$, the averaging uses the conditional probability $P(\cdot|\mathcal{D})$.

Compute $P(X = x|Y = c, \mathcal{D})$

- The integrals in the following are a short hand for a multiple integral.

$$
\begin{aligned}
& P(X = x|Y = c, \mathcal{D}) \\
= & \int P(X = x|Y = c, \theta, \mathcal{D}) P(\theta|Y = c, \mathcal{D}) d\theta \\
\text{(see later)} = & \int P(X = x|Y = c, \theta) P(\theta|\mathcal{D}) d\theta \\
\text{(Assumption 2 and (7))} = & \int \prod_{j=1}^{D} P(X_j = x_j|Y = c, \theta_{jc}) P(\theta_{jc}|\mathcal{D}) d\theta_{jc} \\
= & \prod_{j=1}^{D} \int P(X_j = x_j|Y = c, \theta_{jc}) P(\theta_{jc}|\mathcal{D}) d\theta_{jc}
\end{aligned}
$$

- In the derivation, we used (prove this)

$$P(X = x|Y = c, \theta, \mathcal{D}) = P(X = x|Y = c, \theta).$$

  The proof is similar to that of (19).

- We also used (prove this)

$$P(\theta|Y = c, \mathcal{D}) = P(\theta|\mathcal{D}).$$

- Recall $P(\theta_{jc}|\mathcal{D})$ is a beta distribution Beta$(N_{jc} + \beta_0, N_c - N_{jc} + \beta_1)$.

- Consider each $\int P(X_j = x_j|Y = c, \theta_{jc})P(\theta_{jc}|\mathcal{D})d\theta_{jc}$.

When $x_j = 1$:

$$\begin{aligned}
&\int P(X_j = x_j | Y = c, \theta_{jc}) P(\theta_{jc}|\mathcal{D}) d\theta_{jc} \\
=&\int \theta_{jc} P(\theta_{jc}|\mathcal{D}) d\theta_{jc} \\
=&E[\Theta_{jc}|\mathcal{D}] = \frac{N_{jc} + \beta_0}{N_c + \beta_0 + \beta_1}.
\end{aligned}$$

When $x_j = 0$:

$$\begin{aligned}
&\int P(X_j = x_j | Y = c, \theta_{jc}) P(\theta_{jc}|\mathcal{D}) d\theta_{jc} \\
=&\int (1 - \theta_{jc}) P(\theta_{jc}|\mathcal{D}) d\theta_{jc} \\
=&1 - E[\Theta_{jc}|\mathcal{D}] = \frac{N_c - N_{jc} + \beta_1}{N_c + \beta_0 + \beta_1}.
\end{aligned}$$

- For shorter notation, let $\bar{\theta}_{jc} \triangleq E[\Theta_{jc}|\mathcal{D}]$.

- The two results can be written together as

$$\bar{\theta}_{jc}^{\mathbf{1}(x_j=1)}(1-\bar{\theta}_{jc})^{\mathbf{1}(x_j=0)},$$

which is the same as

$$\bar{\theta}_{jc}^{x_j}(1-\bar{\theta}_{jc})^{1-x_j}.$$

- We then have

$$P(X=x|Y=c,\mathcal{D})=\prod_{j=1}^{D}\bar{\theta}_{jc}^{x_j}(1-\bar{\theta}_{jc})^{1-x_j}. \tag{21}$$

- This is understandable because $P(X_j=1|Y=c,\Theta_{jc})=\Theta_{jc}$. When computing $P(X_j=1|Y=c,\mathcal{D})$, we are given $\mathcal{D}$ but not Θ_{jc}. we must average out Θ_{jc} over $P(\theta_{jc}|\mathcal{D})$.

Posterior Predictive Distribution - Conclusion

- Putting (20) and (21) into (18), the predictive probability satisfies the following:

$$P(Y = c|X = x, \mathcal{D}) \propto \bar{\pi}_c \prod_{j=1}^{D} \bar{\theta}_{jc}^{x_j} (1 - \bar{\theta}_{jc})^{1-x_j}. \tag{22}$$

  We see the right hand side has the same form as the expression in (17). The difference is that $\hat{\pi}_{MAP}$ and $\hat{\theta}_{MAP}$ are used in (17).

- To classify the document with feature vector x, we need to compute $\log\left(\bar{\pi}_c \prod_{j=1}^{D} \bar{\theta}_{jc}^{x_j} (1 - \bar{\theta}_{jc})^{1-x_j}\right)$ for each c. We then assign the document to a class $\hat{c}$ that maximizes the expression. That is,

$$\hat{c} \in \operatorname*{argmax}_c \left(\log \bar{\pi}_c + \sum_{j=1}^{D} \left(x_j \log \bar{\theta}_{jc} + (1 - x_j) \log(1 - \bar{\theta}_{jc})\right) \right).$$

Proof: $P(X = x|Y = c, \theta, \mathcal{D}) = P(X = x|Y = c, \theta)$

- Intuition: X depends on the random variable Θ. Given $\Theta = \theta$, the training data provides no further information.

Proof:

$$
\begin{aligned}
& P(X=x|Y=c,\theta,\mathcal{D}) \\
=& \frac{P(X=x,Y=c,\mathcal{D}|\theta)}{P(Y=c,\mathcal{D}|\theta)} \\
=& \int \frac{P(X=x,Y=c,\mathcal{D}|\theta,\pi)}{P(Y=c,\mathcal{D}|\theta)} P(\pi|\theta)d\pi \\
\text{(model Assumption 3)} =& \int \frac{P(X=x,Y=c|\theta,\pi)P(\mathcal{D}|\theta,\pi)}{P(Y=c,\mathcal{D}|\theta)} P(\pi|\theta)d\pi \\
=& \int \frac{P(X=x|Y=c,\pi,\theta)P(Y=c|\pi,\theta)P(\mathcal{D}|\theta,\pi)}{P(\mathcal{D},Y=c|\theta)} P(\pi|\theta)d\pi \\
\text{(model Assumption 2)} =& \int \frac{P(X=x|Y=c,\theta)P(Y=c|\pi,\theta)P(\mathcal{D}|\theta,\pi)}{P(\mathcal{D},Y=c|\theta)} P(\pi|\theta)d\pi \\
\text{(model Assumption 3)} =& \int \frac{P(X=x|Y=c,\theta)P(Y=c,\mathcal{D}|\theta,\pi)}{P(\mathcal{D},Y=c|\theta)} P(\pi|\theta)d\pi \\
=& \frac{P(X=x|Y=c,\theta)}{P(\mathcal{D},Y=c|\theta)} \int P(Y=c,\mathcal{D}|\theta,\pi)P(\pi|\theta)d\pi \\
=& \frac{P(X=x|Y=c,\theta)}{P(\mathcal{D},Y=c|\theta)} P(\mathcal{D},Y=c|\theta) \\
=& P(X=x|Y=c,\theta).
\end{aligned}
$$

Proof: $P(\theta|Y=c,\mathcal{D}) = P(\theta|\mathcal{D})$

- To see $P(\theta|Y=c,\mathcal{D}) = P(\theta|\mathcal{D})$ intuitively, Θ has to do with the X random variables and Π has to do with the Y random variable. Knowing $Y=c$ does not tell anything about Θ.

- First, we show

$$P(Y=c|\pi,\theta,\mathcal{D}) = P(Y=c|\pi,\theta). \tag{23}$$

$$\begin{aligned}
P(Y=c|\pi,\theta,\mathcal{D}) &= \frac{P(Y=c,\mathcal{D}|\pi,\theta)}{P(\mathcal{D}|\pi,\theta)} \\
\text{(model Assumption 3)} &= \frac{P(Y=c|\pi,\theta)P(\mathcal{D}|\pi,\theta)}{P(\mathcal{D}|\pi,\theta)} \\
&= P(Y=c|\pi,\theta).
\end{aligned}$$

- We now complete the proof.

$$
\begin{aligned}
P(\theta|Y=c,\mathcal{D}) &= \int P(\theta,\pi|Y=c,\mathcal{D})d\pi \\
\text{(Bayes')} &= \int \frac{P(Y=c|\theta,\pi,\mathcal{D})P(\theta,\pi|\mathcal{D})}{P(Y=c|\mathcal{D})}d\pi \\
\text{(by (23))} &= \int \frac{P(Y=c|\theta,\pi)P(\theta,\pi|\mathcal{D})}{P(Y=c|\mathcal{D})}d\pi \\
\text{(model Assumption 1)} &= \int \frac{P(Y=c|\pi)P(\theta,\pi|\mathcal{D})}{P(Y=c|\mathcal{D})}d\pi \\
\text{(by (8))} &= \int \frac{P(Y=c|\pi)P(\pi|\mathcal{D})P(\theta|\mathcal{D})}{P(Y=c|\mathcal{D})}d\pi \\
\text{(by (19))} &= \int \frac{P(Y=c|\pi,\mathcal{D})P(\pi|\mathcal{D})P(\theta|\mathcal{D})}{P(Y=c|\mathcal{D})}d\pi \\
&= \frac{P(\theta|\mathcal{D})}{P(Y=c|\mathcal{D})}\int P(Y=c|\pi,\mathcal{D})P(\pi|\mathcal{D})d\pi \\
&= P(\theta|\mathcal{D}).
\end{aligned}
$$

Factored Model

- Suppose there are two (random) parameters Π and Θ, which may be vector-valued.
- Assumptions:

  A1: Suppose the likelihood function factorizes:

  $$P(\mathcal{D}|\pi, \theta) = h(\mathcal{D}, \pi) g(\mathcal{D}, \theta).$$

  Here, $P(\mathcal{D}|\pi, \theta)$ is conditional probability. h and g are just functions.

  A2: Suppose the prior factorizes:

  $$P_{\Pi,\Theta}(\pi, \theta) = P_{\Pi}(\pi) P_{\Theta}(\theta).$$

- We will show the posterior factorizes:

  $$P(\pi, \theta|\mathcal{D}) = P(\pi|\mathcal{D}) P(\theta|\mathcal{D}). \tag{24}$$

Proof:

$$\begin{aligned} P(\pi, \theta | \mathcal{D}) &= \frac{P(\mathcal{D}|\pi, \theta) P(\pi, \theta)}{P(\mathcal{D})} \\ &= P(\pi) h(\mathcal{D}, \pi) \frac{P(\theta) g(\mathcal{D}, \theta)}{P(\mathcal{D})}. \end{aligned}$$

Note that

$$\begin{aligned} P(\mathcal{D}) &= \int_\pi \int_\theta P(\mathcal{D}|\pi, \theta) P(\pi, \theta) d\theta d\pi \\ &= \int_\pi \int_\theta h(\mathcal{D}, \pi) g(\mathcal{D}, \theta) P(\pi) P(\theta) d\theta d\pi \\ &= \left(\int P(\pi) h(\mathcal{D}, \pi) d\pi \right) \left(\int P(\theta) g(\mathcal{D}, \theta) d\theta \right). \end{aligned} \tag{25}$$

Then,

$$P(\pi, \theta | \mathcal{D}) = \frac{P(\pi) h(\mathcal{D}, \pi)}{\int P(\pi) h(\mathcal{D}, \pi) d\pi} \frac{P(\theta) g(\mathcal{D}, \theta)}{\int P(\theta) g(\mathcal{D}, \theta) d\theta}.$$

We then have

$$P(\pi|\mathcal{D}) = \int P(\pi,\theta|\mathcal{D})d\theta$$

$$= \frac{P(\pi)h(\mathcal{D},\pi)}{\int P(\pi)h(\mathcal{D},\pi)d\pi}. \tag{26}$$

Similarly,

$$P(\theta|\mathcal{D}) = \frac{P(\theta)g(\mathcal{D},\theta)}{\int P(\theta)g(\mathcal{D},\theta)d\theta}. \tag{27}$$

We then we get the factorization in (24).

Consequences of Factored Models

- For an MAP estimate, we need to find

$$\operatorname*{argmax}_{\pi,\theta} P(\pi, \theta|\mathcal{D}).$$

The factorization of $P(\pi, \theta|\mathcal{D})$ leads to two sub-problems:

$$\operatorname*{argmax}_{\pi} P(\pi|\mathcal{D}), \qquad \operatorname*{argmax}_{\theta} P(\theta|\mathcal{D}).$$

- Furthermore, since

$$P(\theta|\mathcal{D}) \propto P(\theta)g(\mathcal{D}, \theta),$$

we have

$$\operatorname*{argmax}_{\theta} P(\theta|\mathcal{D}) = \operatorname*{argmax}_{\theta} P(\theta)g(\mathcal{D}, \theta).$$

Similarly,

$$\operatorname*{argmax}_{\pi} P(\pi|\mathcal{D}) = \operatorname*{argmax}_{\pi} P(\pi)h(\mathcal{D}, \pi).$$

Sufficient Statistic

- In (3), the data $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ shows up in the form of N_c and N_{jc} for different j and c.

  When consider the random sample $\{(X_1, Y_1), \ldots, (X_N, Y_N)\}$, the corresponding statistics N_c and N_{jc}, for all j and c, are jointly sufficient statistics.

  General discussion:

- Let $X = (X_1, X_2, \ldots, X_n)$ be a random sample. Suppose the common distribution P_{X_i} depends on the parameter θ (in general, a vector).

  Let $x = (x_1, \ldots, x_n)$.

  Important Note: X and x are defined differently from before.

- A **statistic** is a function $T = r(X)$ of the sample. Examples:

- the sample mean: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $T_1 = \max\{X_1, \ldots, X_n\}$

- Suppose we wish to estimate the unknown parameter θ based on the sample.

  Informally, a statistic $T = r(X)$ is called a sufficient statistic if one can estimate θ based on T just as well as based on the entire sample. Formally,

  Definition 1: A statistic $T = r(X)$ is called a **sufficient statistic** for θ if the conditional distribution of X given $T = t$ does not depend on θ.

Why Definition 1?

- For notational simplicity, consider the discrete case.
- Suppose every time the random sample X takes the value x, one is given $r(x)$ instead of x.
- Since T is a sufficient statistic, the conditional probability $P(X = x|T = t)$ does not depend on θ and it can be computed. To see that, we have

$$P(X = x|T = t) = \frac{P(X = x, T = t)}{P(T = t)}.$$

The above is equal to 0 if $t \neq r(x)$. We only need to consider the case where $t = r(x)$. Then,

$$P(X = x|T = t) = \frac{P(X = x, r(X) = r(x))}{P(r(X) = t)} = \frac{P(X = x)}{P(X \in r^{-1}(t))}.$$

- Thus, for each t, one can use the probability model for each X_i to compute the conditional probability $P(X = x|T = t)$ for $x \in r^{-1}(t)$.

  If this is done analytically, θ should cancel out and the conditional probability doesn't have an unknown parameter. If it is done numerically, one can set θ to be any allowed value.

- Then, when given $t = r(x)$, one can draw a random sample $Y = (Y_1, \ldots, Y_n)$ from the conditional distribution $P(X = x|T = t)$. In other words,

$$P(Y = x|T = t) = P(X = x|T = t).$$

  Then, we must have $P(Y = x) = P(X = x)$. That is, Y has the same distribution as the random sample X. To see this:

$$\begin{aligned} P(Y = x) =& P(Y = x, r(X) = r(x)) \\ =& P(Y = x | T = t) P(T = t) \\ =& P(X = x | T = t) P(T = t) \\ =& P(X = x, r(X) = r(x)) \\ =& P(X = x). \end{aligned}$$

- Thus, for any estimator of θ using the sample X, say $\hat{\theta}(X)$, one has the estimator $\hat{\theta}(Y)$. The two estimators have the same statistical properties, i.e., they have the same distribution.
- We have shown that knowing $r(X)$ and knowing $X_1, \ldots, X_n$ are really the same for the purpose of estimating θ.

How to check a statistic is sufficient?

- The following theorem can be used to check if a statistic is sufficient.

  Factorization Theorem: Let $f(x|\theta)$ be the joint pdf or pmf. A statistic $T = r(X)$ is sufficient if and only if there are non-negative functions h and g such that

  $$f(x|\theta) = h(x)g(r(x), \theta).$$

  Note that h does not depend on θ; g depends on the sample $X = (X_1, \dots, X_n)$ only through the statistics $r(X)$. h may be a constant.

- Example: Each X_i is uniformly distributed on $[0, \theta]$, where θ is unknown. The joint density is

  $$f(x|\theta) = \theta^{-n}, x_i \in [0, \theta], \ \forall i,$$

  and it is zero elsewhere. We only need to consider the region where

$x_i \geq 0$ for all i. On that region, the density function can be written as:

$$f(x|\theta) = \theta^{-n}\mathbf{1}(x_i \leq \theta,\ \forall i) = \theta^{-n}\mathbf{1}(\max\{x_1, \ldots, x_n\} \leq \theta).$$

By the factorization theorem, $T = \max\{X_1, \ldots, X_n\}$ is a sufficient statistic.

- Since the term $h(x)$ does not depend on θ, we have

$$\operatorname*{argmax}_{\theta} f(x|\theta) = \operatorname*{argmax}_{\theta} g(r(x), \theta).$$

- For MLE, it is enough to keep $r(x)$.
- For two sets of data x and x' with $r(x) = r(x')$, the two sets of MLE are the same under x or x'.

Jointly Sufficient Statistic

- Consider k statistics: $T_i = r_i(X)$, for $i = 1, \ldots, k$.

  The statistics $T_1, \ldots, T_k$ are **jointly sufficient** if for any $t_1, \ldots, t_k$, the conditional distribution of X given $T_1 = t_1, \ldots, T_k = t_k$ does not depend on θ.

  Theorem: Let $X_1, \ldots, X_n$ be a random sample with joint pdf or pmf $f(x|\theta)$. The statistics $T_i = r_i(X)$, where $i = 1, \ldots, k$, are jointly sufficient if and only if there are non-negative functions h and g such that

$$f(x|\theta) = h(x)g(r_1(x), \ldots, r_k(x); \theta).$$

- Example: Consider n IID Gaussian random variables with unknown mean μ and unknown variance σ^2. The joint probability density function is

$$
\begin{aligned}
f(x|\mu,\sigma) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right).
\end{aligned}
$$

Let

$$
r_1(x) = \sum_{i=1}^{n} x_i
$$

$$
r_2(x) = \sum_{i=1}^{n} x_i^2.
$$

We see that $T_1 = \sum_{i=1}^{n} X_i$ and $T_2 = \sum_{i=1}^{n} X_i^2$ are jointly sufficient statistics.

The sample mean and sample variance are also jointly sufficient statistics.

Bayesian Version of Sufficient Statistic

- The parameter is viewed as a random variable, denoted Θ.
- Let $f_{\Theta|X}(\theta|x)$ denote the conditional pdf or conditional pmf of Θ given $X = x$.

  Let $f_{\Theta|T}(\theta|r(x))$ denote the conditional pdf or conditional pmf of Θ given $T = r(X) = r(x)$.

  Definition 2: A statistic $T = r(X)$ is called a **sufficient statistic** if $f_{\Theta|X}(\theta|x) = f_{\Theta|T}(\theta|r(x))$ for any θ and x.

- For MAP estimate, it is enough to keep $r(x)$.

  If we have two data sets $x = (x_1, \ldots, x_n)$ and $x' = (x'_1, \ldots, x'_n)$ with $r(x) = r(x')$, then $f_{\Theta|X}(\theta|x) = f_{\Theta|X}(\theta|x')$. The MAP estimator will yield the same estimate for θ in the two cases.

- Note

$$f_{\Theta|X}(\theta|x) = \frac{f(x|\theta)f_\Theta(\theta)}{f_X(x)},$$

where $f_\Theta(\theta)$ denotes the pdf or pmf of Θ, $f(x|\theta)$ is the likelihood function or joint pdf of $X_1, \ldots, X_n$ conditional on $\Theta = \theta$, and $f_X(x)$ is the unconditioned joint pdf or pmf of $X_1, \ldots, X_n$. Similarly,

$$f_{\Theta|T}(\theta|r(x)) = \frac{f_{T|\Theta}(r(x)|\theta)f_\Theta(\theta)}{f_T(r(x))}.$$

- Then, $f_{\Theta|X}(\theta|x) = f_{\Theta|T}(\theta|r(x))$ implies

$$\frac{f(x|\theta)f_\Theta(\theta)}{f_X(x)} = \frac{f_{T|\Theta}(r(x)|\theta)f_\Theta(\theta)}{f_T(r(x))}.$$

Or,

$$f(x|\theta) = \frac{f_X(x)}{f_T(r(x))} f_{T|\Theta}(r(x)|\theta).$$

- By the factorization theorem, $r(X)$ is a sufficient statistic according to the first definition earlier.

- Since the term $\frac{f_X(x)}{f_T(r(x))}$ does not depend on θ, we have

$$\operatorname*{argmax}_{\theta} f(x|\theta) = \operatorname*{argmax}_{\theta} f_{T|\Theta}(r(x)|\theta).$$

- For MLE, it is enough to keep $r(x)$.