# 1 Why

# 2 How

## 2.1 Notation

## 2.2 Bayesian Approach

Let $n \in \mathbb{N}^+$ the number of relavent features of making a good cup of coffee, e.g. the temperature of water; $x \in \mathbb{R}^n$ the values of the features. Let $Y$ the taste of coffee under a given $x$, which is either 0 (tastes bad) or 1 (tastes good), thus naturally is a random variable obeys a Bernoulli distribution with probability (confidence) $\psi$, i.e. $Y \sim \mathrm{Ber}(\psi)$. Let $f$ the model relates $x$ and $\psi$, depending also on paramters $w \in \mathbb{R}^m$ for some $m \in \mathbb{N}^+$, i.e. $\psi = f(x; w)$.

The Bayesian approach is as follow.

**Theorem 1.** *We have*

$$p(Y=1|X=x) = \mathbb{E}_{w_{(s)} \sim p(W)}[f(x, w_{(s)})],$$

*where $w_{(s)} \sim p(W)$ means that $\{w_{(s)}: s = 1, 2, ...\}$ are sampled from $P(W)$.*

**Proof.** By Bayesian formula,

$$p(Y=1|X=x) = \frac{p(X=x, Y=1)}{p(X=x)}.$$

Then by total probability formula,

$$p(X=x, Y=1) = \int_{\mathbb{R}^m} \mathrm{d}w p(X=x, Y=1, W=w),$$

then Bayesian formula gives

$$p(X=x, Y=1) = \int_{\mathbb{R}^m} \mathrm{d}w p(Y=1|X=x, W=w)\, p(X=x, W=w).$$

Since $x$ and $w$ are independent, $p(X=x, W=w) = p(X=x)\, p(W=w)$. Put all together,

$$\begin{aligned} p(Y=1|X=x) &= \frac{p(X=x, Y=1)}{p(x)} \\ &= \frac{\int_{\mathbb{R}^m} \mathrm{d}w p(Y=1|X=x, W=w)\, p(X=x, W=w)}{p(X=x)} \\ &= \frac{\int_{\mathbb{R}^m} \mathrm{d}w p(Y=1|X=x, W=w)\, p(X=x)\, p(W=w)}{p(X=x)} \\ &= \int_{\mathbb{R}^m} \mathrm{d}w p(Y=1|X=x, W=w)\, p(W=w), \end{aligned}$$

or simply,

$$p(Y=1|X=x) = \mathbb{E}_{w_{(s)}}[p(Y=1|X=x, W=w_{(s)})].$$

And then insert $f$ as

$$p(Y=1|X=x, W=w) = f(x, w),$$

since $Y \sim \mathrm{Ber}(f(x, w))$. So, in one word,

$$p(Y=1|X=x) = \mathbb{E}_{w_{(s)} \sim p(W)}[\, f(x, w_{(s)})]. \qquad \square$$

What we want to find is a $x_*$, s.t. $p(Y=1|X=x)$ $(=\mathbb{E}_{w_{(s)} \sim p(W)}[\, f(x, w_{(s)})])$ is maximized. Or say, we are searching

$$x_* = \underset{x}{\operatorname{argmax}} \{\mathbb{E}_{w_{(s)} \sim p(W)}[\, f(x, w_{(s)})]\}.$$

However, the only thing we have not known yet is the distribution of $W$. We are so humble that know nothing on how to make a good cup of coffee, so we use a flatten prior of $W$, i.e. $W \sim \mathrm{Uniform}$, with some support wide enough. We can obtain the posterior of $W$ by inserting the data, i.e. a list of pairs $(x, y)$, as the value of $Y$ (the taste of a cup of coffee) given by some $x$. By feeding the data, we iterative gain the prior of $W$, which is the posterior in the previous iteration, as

$$p_{i+1}(W=w) = \frac{p(Y=y_i, X=x_i|W=w)\, p_i(W=w)}{p(Y=y_i, X=x_i)}.$$

**Theorem 2.** *If define $g(a, b)$ as $b$ if $a=1$ and as $1-b$ if $a=0$, and if initially use flatten prior, i.e. $p_1 = \mathrm{Const}$, then we have, for data $D := \{x_i, y_i : i = 1, 2, ..., N\}$,*

$$p_N(W=w) = c(D) \times \prod_{i=1}^{N} g(y_i, f(x_i, w)),$$

*or say,*

$$\ln[p_N(W=w)] = \sum_{i=1}^{N} \ln[g(y_i, f(x_i, w))] + \ln[c(D)],$$

*where $c(D)$ can also be seen as the normalization factor of $p_N(W=w)$ since Bayesian formula always ensures normalization of probability.*

**Proof.** By Bayesian formula and the independence between $X$ and $W$,

$$p(Y=y_i, X=x_i|W=w) = p(Y=y_i|X=x_i, W=w)\, p(X=x_i)$$

and

$$p(Y=y_i, X=x_i) = p(Y=y_i|X=x_i)\, p(X=x_i),$$

thus

$$p_{i+1}(W=w) = p_i(W=w)\, \frac{p(Y=y_i|X=x_i, W=w)}{p(Y=y_i|X=x_i)};$$

and since we have known in previous that $p(Y=1|X=x_i) = \mathbb{E}_{w_{(s)} \sim p_i(W)}[\, f(x_i, w_{(s)})]$ and likewise $p(Y=0|X=x_i) = \mathbb{E}_{w_{(s)} \sim p_i(W)}[1 - f(x_i, w_{(s)})]$, we finally get, if $y_i = 1$

$$p_{i+1}(W=w) = p_i(W=w)\, \frac{f(x_i, w)}{\mathbb{E}_{w_{(s)} \sim p_i(W)}[\, f(x_i, w_{(s)})]},$$

else $(y_i = 0)$

$$p_{i+1}(W = w) = p_i(W = w)\,\frac{1 - f(x_i, w)}{\mathbb{E}_{w_{(s)}\sim p_i(W)}[\,1 - f(x_i, w_{(s)})]}.$$

After the first iteration, by $x_1$ and $y_1 = 1$,

$$p_2(W = w) = \text{Const}\,\frac{f(x_1, w)}{\mathbb{E}_{w_{(s)}\sim \text{Uniform}}[\,f(x_1, w_{(s)})]} = c(x_1, y_1)\, f(x_1, w).$$

Then the next iteration, suppose $y_2 = 1$ still,

$$p_3(W = w) = \{c(x_1, y_1)\, f(x_1, w)\}\left\{\frac{f(x_2, w)}{\mathbb{E}_{w_{(s)}\sim p_2(W)}[\,f(x_2, w_{(s)})]}\right\},$$

and re-define $c(\{(x_1, y_1), (x_2, y_2)\}) := c(x_1, y_1)/\mathbb{E}_{w_{(s)}\sim p_2(W)}[\,f(x_2, w_{(s)})]$, thus

$$p_3(W = w) = c(\{(x_1, y_1), (x_2, y_2)\})\, f(x_1, w)\, f(x_2, w).$$

And if $y_2 = 0$,

$$p_3(W = w) = c(\{(x_1, y_1), (x_2, y_2)\})\, f(x_1, w)[\,1 - f(x_2, w)].$$

So, generally, if define $g(a, b)$ as $b$ if $a = 1$ and as $1 - b$ if $a = 0$, then for data $D := \{x_i, y_i\colon i = 1, 2, ..., N\}$,

$$p_N(W = w) = c(D) \times \prod_{i=1}^{N} g(y_i, f(x_i, w)),$$

or say,

$$\ln[p_N(W = w)] = \sum_{i=1}^{N} \ln[g(y_i, f(x_i, w))] + c(D).$$ □

**Corollary 3.** *If data* $D = \{x_{\text{BEST}}, y_i = 1\colon i = 1, 2, ..., N\}$, *then*

$$\lim_{N\to+\infty} \operatorname*{argmax}_x \big\{\mathbb{E}_{w_{(s)}\sim p_N(W)}[f(x, w_{(s)})]\big\} = x_{\text{BEST}}.$$

**Proof.** XXX $p_N(W = w) = c(D)\,[f(x_{\text{BEST}}, w)]^N$ □

**Algorithm 1**

XXX (init)

1. $D \leftarrow D \cup (x_i, y_i)$;
2. $\ln[p(W = w)] \leftarrow \ln[p(W = w)] + \ln[g(y_i, f(x_i, w))]$;
3. fit $p(W)$ by variational inference;
4. sample $\{w_s\colon s = 1, 2, ..., N_s\}$ from $p(W)$;
5. $x_* = \operatorname*{argmax}_x \{\mathbb{E}_{w_s}[f(x, w_s)]\}$;
6. Make a cup of coffee by feature values $x_*$;
7. Taste the cupe of coffee;
8. Return your opinion as $y_*$;
9. $(x_{i+1}, y_{i+1}) \leftarrow (x_*, y_*)$.

## 2.3 An Instant Model