Introduction to Machine Learning
Lecture 2 - MAP Estimator

1 Maximum A Posteriori (MAP) Classifier (Estimator)

1.1 Example

Consider two fair coins W_1 and W_2:

$$W_{1,2} = \begin{cases} 0 & \text{w.p. } \frac{1}{2} \\ 1 & \text{w.p. } \frac{1}{2} \end{cases}$$

$$W_1, W_2 \in \Omega = \{0, 1\}$$

We denote their sum by:

$$X = W_1 + W_2$$

$$X \in \mathcal{X} = \{0, 1, 2\}$$

Question A toss was made and we are given with the sum $X = W_1 + W_2$.
Given $X = x_0$ we want to estimate the value of W_1.
Find an estimator function $\hat{W} : \mathcal{X} \longrightarrow \Omega$ which minimize the probability of error:

$$\boxed{\hat{W}_1(x_0) = \arg\min_{W \in \Omega} \Pr\{W_1 \neq W | X = x_0\} = ?}$$

Solution:
For every possible value of $x_0 \in \mathcal{X}$,
we will calculate the value $\Pr(W_1 \neq W | X = x_0)$, once for $W = 0$, and once for $W = 1$.

1. For $x_0 = 0$ we have:

$$\begin{cases} \Pr\{W_1 \neq 0 | X = 0\} = 1 - \Pr\{W_1 = 0 | X = 0\} = 1 - 1 = 0 \\ \Pr\{W_1 \neq 1 | X = 0\} = 1 - \Pr\{W_1 = 1 | X = 0\} = 1 - 0 = 1 \end{cases}$$

$$\Rightarrow \Pr\{W_1 \neq 0 | X = 0\} < \Pr\{W_1 \neq 1 | X = 0\} \Rightarrow \boxed{\hat{W}_1(0) = 0}$$

2. For $x_0 = 2$ we have:

$$\begin{cases} \Pr\{W_1 \neq 0 | X = 2\} = 1 - \Pr\{W_1 = 0 | X = 2\} = 1 - 0 = 1 \\ \Pr\{W_1 \neq 1 | X = 2\} = 1 - \Pr\{W_1 = 1 | X = 2\} = 1 - 1 = 0 \end{cases}$$

$$\Rightarrow \Pr\{W_1 \neq 1 | X = 2\} < \Pr\{W_1 \neq 0 | X = 2\} \Rightarrow \boxed{\hat{W}_1(2) = 1}$$

3. For $x_0 = 1$ we have:

$$\begin{cases} \Pr\{W_1 \neq 0 | X = 1\} = 1 - \Pr\{W_1 = 0 | X = 1\} = \frac{1}{2} \\ \Pr\{W_1 \neq 1 | X = 1\} = 1 - \Pr\{W_1 = 1 | X = 1\} = \frac{1}{2} \end{cases}$$

$$\Rightarrow \boxed{\Pr\{W_1 \neq 0 | X = 1\} = \Pr\{W_1 \neq 1 | X = 1\} = \frac{1}{2}}$$

Thus, for $x_0 = 1$, the probability of error is the same for any choice of $\hat{W}_1(1) \in \Omega$.

Overall, the estimator is given by:

$$\Rightarrow \boxed{\hat{W}_1(x_0) = \begin{cases} 0 & x = 0 \\ 0 & x = 1 \\ 1 & x = 2 \end{cases}} \text{(Or equivalently } \hat{W}_1(1) = 1)$$

1.2 Formulation

Let Ω be the set of all possible classes (states):

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_C\}$$

Let X be a random variable (which depends on the state ω) with conditional probability function $p_{X|\Omega}$.

1.2.1 A priori probability

The a priori probability $P_\Omega(\omega_i) = \Pr\{\omega = \omega_i\}$ describes the probability that the state ω is ω_i:

$$P_\Omega : \Omega \longrightarrow [0,1]$$

1.2.2 Conditional probability

The conditional probability $p_{X|\Omega}(x|\omega)$ (or just $p(x|\omega)$) describes the probability of X given ω (that is, ω is known).
By definition:

$$\boxed{p_{X|\Omega}(x|\omega) \triangleq \frac{p_{X,\Omega}(x,\omega)}{P_\Omega(\omega)}}$$

Notes

1. $p(x|\omega)$ is in fact, the likelihood function $\mathcal{L}(\omega)$ (for a fixed x).
2. $p(x,\omega) = p(x|\omega) \cdot P(\omega)$

1.3 Classifier performance measurement

Assume that we have some pair $(\boldsymbol{x}_0, \omega_0)$, namely, the input $\boldsymbol{x}_0 \in \mathcal{X}$ originated from the state $\omega_0 \in \Omega$.
Ideally, we want to find a classifier $\hat{\omega} : \mathcal{X} \longrightarrow \Omega$ such that for any pair $(\boldsymbol{x}_0, \omega_0)$:

$$\hat{\omega}(\boldsymbol{x}_0) = \omega_0$$

Note that for some pairs, this task is impossible.

Example Let

$$\Omega = \{\text{male, female}\}$$

Let $\mathcal{X} = \mathbb{R}^+$ be the height of a grown person.
Thus (approximately):

$$x|\omega \sim \begin{cases} \mathcal{N}(165, 10^2) & \omega = \text{male} \\ \mathcal{N}(155, 8^2) & \omega = \text{female} \end{cases}$$

Note that (intuitively):

1. $\hat{\omega}(200) = \text{male}$, (with high confidence)
2. $\hat{\omega}(150) = \text{female}$, (with high confidence)
3. $\hat{\omega}(160) = ?$ (hard to deice, weak confidence)

We will try to find a classifier which minimize the probability of error.

1.3.1 Probability of error

The conditional error of the classifier $\hat{\omega}$ on the input $\boldsymbol{x}_0$ is given by:

$$\boxed{\Pr\left\{\text{error}|X=\boldsymbol{x}_0\right\} \triangleq \Pr\left\{\omega \neq \hat{\omega}\left(\boldsymbol{x}_0\right)|X=\boldsymbol{x}_0\right\} = 1-\Pr\left\{\omega = \hat{\omega}\left(\boldsymbol{x}_0\right)|X=\boldsymbol{x}_0\right\} = 1-P_{\Omega|X}\left(\hat{\omega}\left(\boldsymbol{x}_0\right)|\boldsymbol{x}_0\right)}$$

The probability of error is given by:

$$\begin{aligned}\Pr\left\{\text{error}\right\} \triangleq \mathbb{E}\left[\Pr\left\{\text{error}|X\right\}\right] &= \int_{\mathcal{X}} \Pr\left\{\text{error}|X=\boldsymbol{x}\right\} p_X\left(\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\ &= \int_{\mathcal{X}} \left(1-P_{\Omega|\mathcal{X}}\left(\hat{\omega}\left(\boldsymbol{x}\right)|\boldsymbol{x}\right)\right) p_X\left(\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\\ &= 1-\int_{\mathcal{X}} P_{\Omega|X}\left(\hat{\omega}\left(\boldsymbol{x}\right)|\boldsymbol{x}\right) p_X\left(\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}\end{aligned}$$

1.3.2 The optimal (MAP) classifier

Consider the conditional error:

$$\Pr\left\{\text{error}|X=\boldsymbol{x}\right\} = 1-P_{\Omega|X}\left(\hat{\omega}\left(\boldsymbol{x}\right)|\boldsymbol{x}\right)$$

Note that $\Pr\left\{\text{error}|X=\boldsymbol{x}\right\}$ attains its minimum value when $P_{\Omega|X}\left(\hat{\omega}\left(\boldsymbol{x}\right)|\boldsymbol{x}\right)$ achieves its maximum value.
In other words:

$$\arg\min_{\omega} \Pr\left\{\text{error}|X=\boldsymbol{x}\right\} = \arg\min_{\omega} 1-P_{\Omega|X}\left(\omega|\boldsymbol{x}\right) = \arg\max_{\omega} P_{\Omega|X}\left(\omega|\boldsymbol{x}\right)$$

So we denote:

$$\boxed{\hat{\omega}_{MAP}\left(\boldsymbol{x}\right) \triangleq \arg\max_{\omega} P_{\Omega|X}\left(\omega|\boldsymbol{x}\right)}$$

$P_{\Omega|X}\left(\omega|\boldsymbol{x}\right)$ is known as the a posteriori probability.
Namely, $\hat{\omega}_{MAP}$ is the MAP (Maximum A Posteriori) classifier (estimator).

Theorem 1. *$\hat{\omega}_{MAP}$ also minimize the probability of error $Pr\left(error\right)$.*

Proof. The probability of error is given by:

$$\Pr\left\{\text{error}\right\} = 1-\int_{\mathcal{X}} P\left(\hat{\omega}\left(\boldsymbol{x}\right)|\boldsymbol{x}\right) p_X\left(\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}$$

To minimize the error term one should maximize the integral value.
Since $p_X\left(\boldsymbol{x}\right) \geq 0$ the integral will obtain its maximum value for $\hat{\omega}\left(\boldsymbol{x}\right)$ such that $P_{\Omega|X}\left(\hat{\omega}\left(\boldsymbol{x}\right)|\boldsymbol{x}\right)$ is maximized.
By definition, $P_{\Omega|X}\left(\omega|\boldsymbol{x}\right)$ is maximized by setting $\hat{\omega}\left(\boldsymbol{x}\right) = \hat{\omega}_{MAP}\left(\boldsymbol{x}\right)$. □

1.3.3 Bayes' Law:

Usually, we only know the a priori probability P_{Ω} and the conditional probability $p_{X|\Omega}$.

Question How can we compute:

$$\hat{\omega}_{MAP}\left(\boldsymbol{x}\right) = \arg\max_{\omega} P_{\Omega|X}\left(\omega|\boldsymbol{x}\right) = ?$$

without knowing the a postriori probability: $P_{\Omega|X}\left(\omega|\boldsymbol{x}\right)$.

Solution: By using Bayes' law.
Remember that, by definition:

$$P(A|B) \triangleq \frac{P(A,B)}{P(B)}$$

Thus, Bayes' law state that:

$$\Rightarrow \boxed{p(\omega|\boldsymbol{x}) \triangleq \frac{p(\boldsymbol{x},\omega)}{p_X(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\omega)\, P_\Omega(\omega)}{p_X(\boldsymbol{x})}}$$

Now, using Bayes' law we have:

$$\hat{\omega}_{MAP}(\boldsymbol{x}) = \arg\max_{\omega} P_{\Omega|X}(\omega|\boldsymbol{x}) = \arg\max_{\omega} \frac{p(\boldsymbol{x}|\omega)\, P_\Omega(\omega)}{p_X(\boldsymbol{x})} = \arg\max_{\omega} p(\boldsymbol{x}|\omega)\, P_\Omega(\omega)$$

Where the last equation is true since $\arg\max_\omega$ is independent of the values of $p_X(\boldsymbol{x}) > 0$

$$\Rightarrow \boxed{\hat{\omega}_{MAP}(\boldsymbol{x}) \triangleq \arg\max_{\omega} P_{\Omega|X}(\omega|\boldsymbol{x}) = \arg\max_{\omega} p(\boldsymbol{x}|\omega)\, P_\Omega(\omega)}$$

1.4 Estimation example

Consider the unknown "state" (parameter) $\omega \in \mathbb{R}$, with the following a priori probability:

$$p_{\Omega}(\omega) = \frac{1}{\sqrt{2\pi\sigma_{\omega}^2}} e^{-\frac{(\omega-10)^2}{2\sigma_{\omega}^2}}$$

Given the parameter ω the conditional probability of the random variable X, is given by:

$$p_{X|\Omega}(x|\omega) = \frac{1}{\sqrt{2\pi\sigma_{x}^2}} e^{-\frac{(x-\omega)^2}{2\sigma_{x}^2}}$$

σ_{ω}^2 are σ_x^2 are known and let $\beta \triangleq \frac{\sigma_{\omega}^2}{\sigma_x^2}$.
$\{x_i\}_{i=1}^N$ are N i.i.d realizations generated from the state ω.

- Find the MAP estimator $\hat{\omega}_{MAP}$.

Solution:

$$\boxed{\hat{\omega}_{MAP} = \arg\max_{\omega} p(\omega|\{x_i\}) = \arg\max_{\omega} p(\{x_i\}|\omega)\, p_{\Omega}(\omega)}$$

Since the observations are independent we have:

$$p(\{x_i\}|\omega) = p(x_1, x_2, \ldots, x_N|\omega) = \prod_{i=1}^{N} p(x_i|\omega) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{(x_i-\omega)^2}{2}} = \left(\frac{1}{\sqrt{2\pi\sigma_x^2}}\right)^N e^{-\sum_{i=1}^{N}\frac{(x_i-\omega)^2}{2\sigma_x^2}}$$

$$\begin{aligned}
\Rightarrow \hat{\omega}_{MAP} &= \arg\max_{\omega} p(\{x_i\}|\omega)\, p(\omega) \\
&= \arg\max_{\omega} p(\{x_i\}|\omega) \frac{1}{\sqrt{2\pi\sigma_{\omega}^2}} e^{-\frac{(\omega-10)^2}{2\sigma_{\omega}^2}} \\
&= \arg\max_{\omega} e^{-\sum_{i=1}^{N}\frac{(x_i-\omega)^2}{2\sigma_x^2}} e^{-\frac{(\omega-10)^2}{2\sigma_{\omega}^2}} \\
&= \arg\,\mathbf{min}_{\omega} \sum_{i=1}^{N} \frac{(x_i-\omega)^2}{2\sigma_x^2} + \frac{(\omega-10)^2}{2\sigma_{\omega}^2} \\
&= \arg\min_{\omega} \underbrace{\frac{\sigma_{\omega}^2}{\sigma_x^2}\sum_{i=1}^{N}(\omega-x_i)^2 + (\omega-10)^2}_{\triangleq f(\omega)}
\end{aligned}$$

Let us find the point ω where f obtains its minimum:

$$f'(\omega) = 0$$

$$\underbrace{\frac{\sigma_{\omega}^2}{\sigma_x^2}}_{\triangleq\beta}\sum_{i=1}^{N}(\omega-x_i) + (\omega-10) = 0$$

$$\beta N(\omega-\overline{x}) + \omega - 10 = 0, \qquad \left(\overline{x} \triangleq \frac{1}{N}\sum_{i=1}^{N} x_i\right)$$

$$(\beta N + 1)\,\omega = 10 + \overline{x}$$

$$\omega = \frac{10 + \beta N\overline{x}}{\beta N + 1}$$

$\overline{x} \triangleq \frac{1}{N}\sum_{i=1}^{N} x_i$ is the empirical mean of $\{x_i\}_{i=1}^N$.

Thus:

$$\Rightarrow \boxed{\hat{\omega}_{MAP} = \arg\min_{\omega} f(\omega) = \frac{10 + \beta N \overline{x}}{\beta N + 1}}$$

Notes ($\beta \triangleq \frac{\sigma_\omega^2}{\sigma_x^2}$):

1. If $\beta \to 0$ the estimation is based on the prior knowledge and not the observations:

$$\hat{\omega}_{MAP} \underset{\beta \to 0}{\longrightarrow} 10 = \mathbb{E}[\omega]$$

2. If $\beta \to \infty$ the estimation is based on the observations and not the prior knowledge:

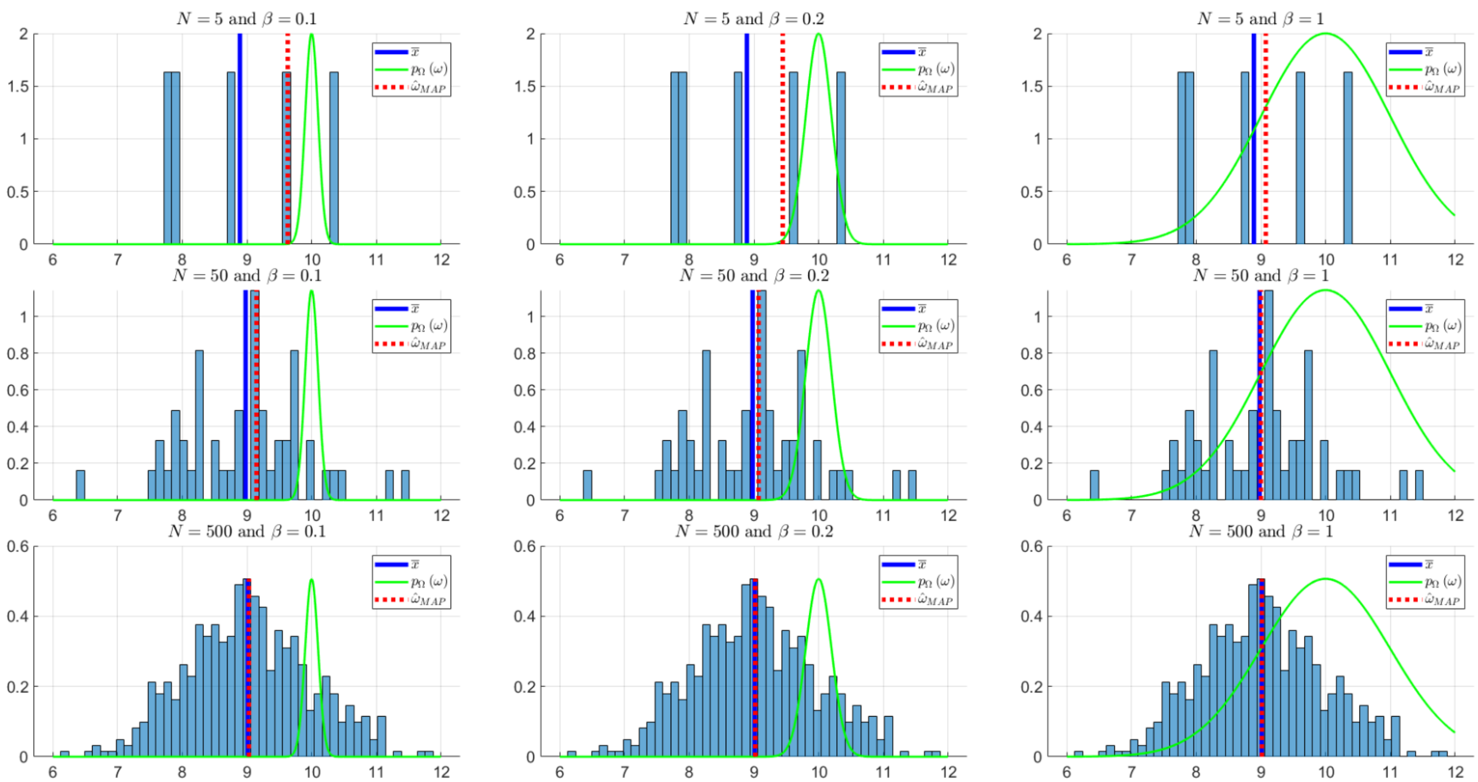
$$\hat{\omega}_{MAP} \underset{\beta \to \infty}{\longrightarrow} \overline{x} = \mathbb{E}[X]$$

3. If $N \to \infty$ the estimation is based on the observations and not the prior knowledge:

$$\hat{\omega}_{MAP} \underset{N \to \infty}{\longrightarrow} \overline{x} = \mathbb{E}[X]$$

Simulation

We test the MAP estimator for different values of N and β.
We draw one realization of ω and then generated N realizations from X (given ω).
Notice the trade-off between the prior knowledge p_Ω and the empirical estimation $\overline{x}$:



1.5 Exercise

Consider the random variable X, such that, given the parameter $\omega \in \mathbb{R}$ we have:

$$p_{X|\Omega}(x|\omega) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\omega)^2}{2}}$$

The prior probability is given by:

$$p_{\Omega}(\omega) = \begin{cases} \frac{1}{2} & \omega \in [-1,1] \\ 0 & \text{else} \end{cases}$$

$\{x_i\}_{i=1}^N$ are N i.i.d realizations generated from the state ω.
Find the MAP estimator $\hat{\omega}_{MAP}$.
Solution:

$$\hat{\omega}_{MAP} = \arg\max_{\omega} p(\omega|\{x_i\}) = \arg\max_{\omega} p(\{x_i\}|\omega) p_{\Omega}(\omega)$$

Since the observations are independent we have:

$$p(\{x_i\}|\omega) = p(x_1, x_2, \ldots, x_N|\omega) = \prod_{i=1}^N p(x_i|\omega) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\omega)^2}{2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^N e^{-\sum_{i=1}^N \frac{(x_i-\omega)^2}{2}}$$

$$\begin{aligned}
\Rightarrow \hat{\omega}_{MAP} &= \arg\max_{\omega} p(\{x_i\}|\omega) p(\omega) \\
&= \arg\max_{\omega\in[-1,1]} p(\{x_i\}|\omega) \\
&= \arg\max_{\omega\in[-1,1]} \left(\frac{1}{\sqrt{2\pi}}\right)^N e^{-\sum_{i=1}^N \frac{(x_i-\omega)^2}{2}} \\
&= \arg\max_{\omega\in[-1,1]} e^{-\sum_{i=1}^N \frac{(x_i-\omega)^2}{2}} \\
&= \arg\min_{\omega\in[-1,1]} \underbrace{\sum_{i=1}^N (x_i-\omega)^2}_{\triangleq f(\omega)}
\end{aligned}$$

We find the minimum of f by comparing the derivative to zeros:

$$f'(\omega) = 0$$

$$-2\sum_{i=1}^N (x_i - \omega) = 0$$

$$\sum_{i=1}^N x_i = N\omega$$

$$\Rightarrow \boxed{\omega = \frac{1}{N}\sum_{i=1}^N x_i \triangleq \overline{x}}$$

where $\overline{x}$ is the empirical mean of $\{x_i\}_{i=1}^N$.
The optimization constrains $\omega \in [-1,1]$ force the solution to that interval.
Since f is an order 2 polynomial (quadratic polynomial), we have:

$$\Rightarrow \boxed{\hat{\omega}_{MAP} = \arg\min_{\omega\in[-1,1]} f(\omega) = \begin{cases} 1 & \overline{x} > 1 \\ \overline{x} & \overline{x} \in [-1,1] \\ -1 & \overline{x} < -1 \end{cases}}$$

1.6 General Loss Function

1.6.1 Loss function

Any function:

$$\ell: \Omega \times \Omega \longrightarrow \mathbb{R}$$

which satisfies:

1. $\ell(\tilde{\omega}, \omega) \geq 0, \qquad \forall \tilde{\omega}, \omega \in \Omega$
2. $\ell(\omega, \omega) = 0, \qquad \forall \omega \in \Omega$

is a loss function.
Using a loss function ℓ implies that some errors are more significant than others.

1.6.2 Risk (weighted error)

Consider some classifier $\hat{\omega}$.

1. The Conditional risk (for a known input $X = \boldsymbol{x}$) is given by:

$$L_x(\boldsymbol{x}) \triangleq \mathbb{E}\left[\ell(\hat{\omega}(\boldsymbol{x}), \omega) \,|\, X = \boldsymbol{x}\right] = \sum_{\omega \in \Omega} \ell(\hat{\omega}(\boldsymbol{x}), \omega) P_{\Omega|X}(\omega|\boldsymbol{x})$$

2. The total risk (averaging all inputs) is given by:

$$L \triangleq \mathbb{E}\left[L_x(X)\right] = \int_{\mathcal{X}} L_x(\boldsymbol{x}) p_X(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

1.6.3 Exercise

Given the following loss function:

$$\ell(\tilde{\omega}, \omega) = \begin{cases} 0 & \tilde{\omega} = \omega \\ C(\omega) & \tilde{\omega} \neq \omega \end{cases}$$

where $C(\omega) > 0$.
Find $\hat{\omega}$ which minimize the total risk L.

$$\hat{\omega} = ?$$

Solution:
we can write:

$$\begin{aligned} L_x(\boldsymbol{x}) &= \sum_{\omega_i \in \Omega} \ell(\hat{\omega}(\boldsymbol{x}), \omega_i) p(\omega_i|\boldsymbol{x}) \\ &\underbrace{=}_{\ell(\omega,\omega)=0} \sum_{\omega_i \neq \hat{\omega}(\boldsymbol{x})} \ell(\hat{\omega}(\boldsymbol{x}), \omega_i) p(\omega_i|\boldsymbol{x}) \\ &= \sum_{\omega_i \neq \hat{\omega}(\boldsymbol{x})} C(\omega_i) p(\omega_i|\boldsymbol{x}) \\ &= \sum_{\omega_i \neq \hat{\omega}(\boldsymbol{x})} C(\omega_i) p(\omega_i|\boldsymbol{x}) + C(\hat{\omega}(\boldsymbol{x})) p(\hat{\omega}(\boldsymbol{x})|\boldsymbol{x}) - C(\hat{\omega}(\boldsymbol{x})) p(\hat{\omega}(\boldsymbol{x})|\boldsymbol{x}) \\ &= \underbrace{\sum_{\omega_i \in \Omega} C(\omega_i) p(\omega_i|\boldsymbol{x})}_{\text{independent of } \hat{\omega}(\boldsymbol{x})} - C(\hat{\omega}(\boldsymbol{x})) p(\hat{\omega}(\boldsymbol{x})|\boldsymbol{x}) \end{aligned}$$

Thus, to minimize $L_x(\boldsymbol{x})$ one should maximize $C(\hat{\omega}(\boldsymbol{x})) p(\hat{\omega}(\boldsymbol{x})|\boldsymbol{x})$. Hence:

$$\Rightarrow \boxed{\hat{\omega}(\boldsymbol{x}) = \arg\max_{\omega} C(\omega) p(\omega|\boldsymbol{x}) = \arg\max_{\omega} C(\omega) P_{\Omega}(\omega) p(\boldsymbol{x}|\omega)}$$

This function $\hat{\omega}$ also minimize L (similar proof as in the original MAP classifier).