Actual Bayesian computation:

Let's now introduce a prior $\mathbb{P}(\theta)$, i.e. a guess about our idea of what is θ.

Recall: MLE: $\underset{\theta}{\operatorname{argmax}} \left(\mathbb{P}(X \mid \theta) \right) = \theta^*_{MLE}$

likelihood of the data

↳ leads to the usual $\hat{\mu} = \frac{1}{N} \sum_i^N x_i$, etc

Now MAP: Maximum a Posteriori:

$$\theta^*_{MAP} = \underset{\theta}{\operatorname{argmax}} \; \mathbb{P}(\theta \mid X) \quad \text{likelihood of the parameters.}$$

$$= \underset{\theta}{\operatorname{argmax}} \left(\frac{\overbrace{\mathbb{P}(X \mid \theta)}^{\text{likelihood of data}} \; \overbrace{\mathbb{P}(\theta)}^{\text{prior}}}{\underbrace{\mathbb{P}(X)}_{\text{evidence (unreachable)}}} \right)$$

$$= \underset{\theta}{\operatorname{argmax}} \left(\underbrace{\log \mathbb{P}(X \mid \theta)}_{\text{as before}} + \underbrace{\log \mathbb{P}(\theta)}_{\text{prior}} \right)$$

Example (simple): $X \sim (x_i)_{1..N}$, $X \sim \mathcal{N}(\underbrace{\mu}_{\theta}, \sigma)$

N examples. Prior: $\mu \sim \mathcal{N}(0, \tau)$.

↳ $\mathbb{P}(\theta) = \frac{1}{\sqrt{2\pi}\,\tau} \cdot e^{-\frac{1}{2} \cdot \frac{\mu^2}{\tau^2}}$

$$\theta^*_{MAP} = \underset{\theta}{\operatorname{argmax}} \left(\log \prod_{i=1}^N \left(\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}} \right) + \log \left(\frac{1}{\sqrt{2\pi}\,\tau} e^{-\frac{1}{2} \frac{\mu^2}{\tau^2}} \right) \right)$$

$$\frac{\delta \ell}{\delta \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^N \left[\log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) - \frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right] + \text{cte} - \frac{1}{2} \frac{\mu^2}{\tau^2}$$

$$= 0 - \frac{\cancel{2}}{\cancel{2}} \sum_i^N \frac{x_i - \mu}{\sigma^2} - \frac{\cancel{2}}{\cancel{2}} \frac{\mu}{\tau^2}$$

$$\iff \sum_{i=1}^N \frac{x_i}{\sigma^2} = N \frac{\mu}{\sigma^2} - \frac{\mu}{\tau^2} \iff \mu \left(1 - \frac{\sigma^2}{N \tau^2} \right) = \frac{1}{N} \sum_i^N x_i$$

So we find: $\hat{\sigma}_{MAP} = \hat{\sigma}_{MLE}$ (here)

$$\hat{\mu}_{MAP} = \hat{\mu}_{MLE} \times \frac{1}{1 - \frac{\sigma^2}{N\tau^2}}$$

(this is just 1 example)

<u>Remarks</u>: $\tau = \infty \Rightarrow$ flat prior $\Rightarrow$ no change

$N \to \infty \Rightarrow \infty$ data $\Rightarrow$ prior becomes irrelevant.

$\sigma \to 0 \Rightarrow$ data little spread $\Rightarrow$ prior irrelevant

We see that the prior acts as a regularizat°

Many interpretations of regul°. as Bayesian priors.

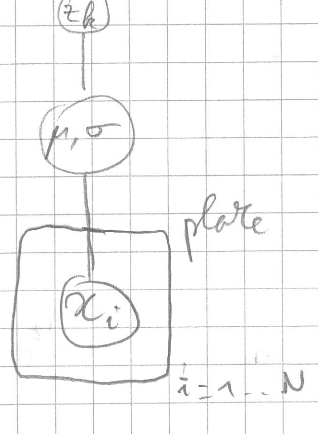
When Sparse distrib° (eg words), need a prior to avoid singularities

N_{train} small $\to$ try to use your knowledge (put it in the prior)

Words distro: Many words $\to$ Not all appear in the train set (corpus): $\hat{\mu}_{word\ \#\ldots} = 0$

But, we expect they can appear (eg test set):

So, need a prior with $p\left(\text{word} = \text{"[some infrequent word]"}\right) > 0$



prior on the prior: assume $\tau \sim \mathcal{N}(m, B)$