

MLE and MAP



Reminder: Bayes' Rule Notation

$$\underbrace{p(w|D)}_{\text{Posterior}} = \frac{\overbrace{p(D|w)}^{\text{Likelihood}} \cdot \overbrace{p(w)}^{\text{Prior}}}{\underbrace{p(D)}_{\text{Marginal Probability of D}}}$$

Reminder: Maximum Likelihood Estimation (MLE)

$$\underbrace{p(w|D)}_{\text{Posterior}} = \frac{\overbrace{p(D|w)}^{\text{Likelihood}} \cdot \overbrace{p(w)}^{\text{Prior}}}{\underbrace{p(D)}_{\text{Marginal Probability of D}}}$$

$$\text{MLE: } \underset{w}{\operatorname{argmax}} p(D|w)$$

- Based on the *likelihood* term
- Find w that makes D the highest probability

Reminder: Maximum A Posteriori (MAP)

$$\underbrace{p(w|D)}_{\text{Posterior}} = \frac{\overbrace{p(D|w)}^{\text{Likelihood}} \cdot \overbrace{p(w)}^{\text{Prior}}}{\underbrace{p(D)}_{\text{Marginal Probability of D}}}$$

$$\text{MAP: } \underset{w}{\operatorname{argmax}} p(w|D) \propto \underset{w}{\operatorname{argmax}} p(D|w)p(w)$$

- Based on the *posterior*
- Find w with the highest probability given D
 - Through Bayes' Rule we find that this is proportional to the likelihood and the prior

MLE vs MAP

$$\text{MLE: } \underset{w}{\operatorname{argmax}} p(D|w)$$

$$\text{MAP: } \underset{w}{\operatorname{argmax}} p(D|w)p(w)$$

- Conceptually different
- Mathematically differ only by multiplication by the prior

Reminder: Negative Log Likelihoods (NLL) (2 min)

- Why is log helpful when optimizing functions?
- Why is it mathematically permissible to take the log of a function when we're maximizing it?
- What do we have to do to a maximization problem to make it mathematically equivalent to minimization?

Reminder: Negative Log Likelihoods (NLL) (2 min)

- Why is log helpful when optimizing functions?
 - Answer: It converts products into sums. Since sums are linear under differentiation, this simplifies computing partial derivatives.
- Why is it mathematically permissible to take the log of a function when we're maximizing it?
 - Answer: The logarithm function is monotonic ($a < b \rightarrow \log(a) < \log(b)$)
- What do we have to do to a maximization problem to make it mathematically equivalent to minimization?
 - Answer: Take the negative i.e. $\operatorname{argmax} f(w) = \operatorname{argmin} -f(w)$

MAP Motivation: Assignment 5

2 MAP Estimation

Rubric: {points:8}

In class, we considered MAP estimation in a regression model where we assumed that:

- The likelihood $p(y_i | x_i, w)$ is a normal distribution with a mean of $w^T x_i$ and a variance of 1.
- The prior for each variable j , $p(w_j)$, is a normal distribution with a mean of zero and a variance of λ^{-1} .

Under these assumptions, we showed that this leads to the standard L2-regularized least squares objective function,

$$f(w) = \frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2,$$

which is the negative log likelihood (NLL) under these assumptions (ignoring an irrelevant constant). For each of the alternate assumptions below, show how the loss function would change (simplifying as much as possible):

MLE Example (4 mins)

The normal distribution notation is $N(\mu, \sigma^2)$

Given $y_i|x_i, w \sim N(w^T x_i, 1)$, which means:

$$p(y_i|x_i, w) = \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp\left(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1}\right)$$

- Given this information:
 - 1) Write the likelihood function for N training examples
 - 2) Use the likelihood function to write an expression for the MLE
 - 3) Solve for the MLE (in matrix notation)

MLE Example: Solution

(1)

$$p(y|X, w) = \prod_{i=1}^N \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp\left(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1}\right) \quad (14)$$

$$\begin{aligned} (2) \quad \arg \max_w p(y|X, w) &= \arg \max_w \prod_{i=1}^N \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp\left(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1}\right) \\ &= \arg \max_w \log\left(\frac{1}{\sqrt{2\pi}}\right) + \sum_{i=1}^N \log\left(\exp\left(-\frac{(w^T x_i - y_i)^2}{2}\right)\right) \quad (\log \text{ is monotonic}) \\ &= \arg \max_w - \sum_{i=1}^N \frac{(w^T x_i - y_i)^2}{2} \\ &= \arg \min_w \frac{1}{2} \cdot \|Xw - y\|_2^2 \quad (\text{negate both sides}) \\ (3) \quad &= \arg \min_w \|Xw - y\|_2^2 \quad (\text{does not change solution}) \end{aligned}$$

MAP Example (4 mins)

The normal distribution notation is $N(\mu, \sigma^2)$

Given $y_i|x_i, w \sim N(w^T x_i, 1)$, which means:

$$p(y_i|x_i, w) = \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp\left(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1}\right)$$

Also assume that our weights are distributed as $w_j \sim N(0, \lambda^{-1})$, i.e.:

$$p(w_j) = \frac{1}{\sqrt{2 \cdot \lambda^{-1} \cdot \pi}} \exp\left(-\frac{(w_j - 0)^2}{2 \cdot \lambda^{-1}}\right)$$

- Given this information:
 - 1) Write the likelihood and prior functions for N training examples
 - 2) Use these functions to write an expression for the MAP
 - 3) Solve for the MAP (in matrix notation)

MAP Example: Solution

(1)

$$p(y|X, w) = \prod_{i=1}^N \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp\left(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1}\right) \quad y_i|x_i, w \sim N(w^T x_i, 1)$$
$$p(w) = \prod_{j=1}^d \frac{1}{\sqrt{2 \cdot \lambda^{-1} \cdot \pi}} \exp\left(-\frac{(w_j - 0)^2}{2 \cdot \lambda^{-1}}\right) \quad w_j \sim N(0, \lambda^{-1})$$

$$\arg \max_w p(w|X, y) = \arg \max_w p(y|X, w) \cdot p(w)$$

(2)

$$= \arg \max_w \log(p(y|X, w)) + \log\left(\prod_{j=1}^d \frac{1}{\sqrt{2 \cdot \lambda^{-1} \cdot \pi}} \exp\left(-\frac{(w_j - 0)^2}{2 \cdot \lambda^{-1}}\right)\right)$$
$$= \arg \max_w \log(p(y|X, w)) + \sum_{j=1}^d \log\left(\exp\left(-\frac{\lambda}{2} w_j^2\right)\right)$$

MAP Example: Solution

$$\arg \max_w p(w|X, y) = \arg \max_w p(y|X, w) \cdot p(w)$$

$$= \arg \max_w \log(p(y|X, w)) + \log\left(\prod_{j=1}^d \frac{1}{\sqrt{2 \cdot \lambda^{-1} \cdot \pi}} \exp\left(-\frac{(w_j - 0)^2}{2 \cdot \lambda^{-1}}\right)\right)$$

$$= \arg \max_w \log(p(y|X, w)) + \sum_{j=1}^d \log\left(\exp\left(-\frac{\lambda}{2} w_j^2\right)\right)$$

$$= \arg \min_w -\log(p(y|X, w)) + \sum_{i=1}^d \frac{\lambda}{2} w_j^2 \quad (\text{negate both sides})$$

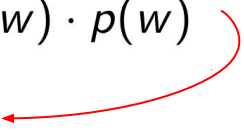
$$(3) \quad = \arg \min_w \frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2$$

Assignment Suggestions

- Follow the same general procedure
 - *Simplify using operations such as applying the log and negative sign to obtain negative log likelihood*
 - *Show the new loss function*
- General form (for MAP), given: $p(y|X, w) = ?$
 $p(w) = ?$
- Want to find:

$$\begin{aligned}\arg \max_w P(w|X, y) &= \arg \max_w p(y|X, w) \cdot p(w) \\ &= \arg \min_w \text{Loss}\end{aligned}$$

Apply log + negative



MAP Example

$$w_j \sim N(0, \lambda^{-1}) \longrightarrow + \frac{\lambda}{2} ||w||^2$$

- MAP loss here differs only by a regularization term
- The regularization term here which we associate with L2 regularization comes from assuming a normal distribution as a prior
- *Bonus*: Is there a prior distribution we can assume that results in a regularization term of 0?