Machine Learning

Supervised Machine Learning – Regression

Part 4: ML and Bayesian regression

The “Bias” Term

Data Model Reminder

- Up until now:

$$y = f(\boldsymbol{x}) + \epsilon$$

For parametric regression, we have some idea on the structure of f (encoded with parameters)

$$E.g., \ f(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})^T \boldsymbol{w}, \text{ for } \boldsymbol{b}(\boldsymbol{x}) = [1, \ \phi_1(\boldsymbol{x}), \dots, \phi_M(\boldsymbol{x})]^T$$

Error statistics: $E\{\epsilon\} = 0$ and $E\{\epsilon^2\} = V$

No error distribution.

"Bias" or "Intercept" Term

$$y = m(\boldsymbol{x};\ \boldsymbol{w}) + \epsilon$$

$$m(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{b}(\boldsymbol{x})^T \boldsymbol{w}, \text{ for } \boldsymbol{b}(\boldsymbol{x}) = [1,\ \phi_1(\boldsymbol{x}), \dots, \phi_M(\boldsymbol{x})]^T$$

$$m(\boldsymbol{x}; \boldsymbol{w}) = \sum_{m=1}^{M} \phi_m(\boldsymbol{x})\, w_{m+1} + w_1$$

$$\text{Train−MSE}(\boldsymbol{w}) = \frac{1}{\mathrm{N_{TR}}} \sum_{\mathrm{n=1}}^{\mathrm{N_{TR}}} \left| y_{TR}(n) - \left(\sum_{m=1}^{M} \phi_m(\boldsymbol{x})\, w_{m+1} + w_1 \right) \right|^2$$

Stationarity condition: Train-MSE is minimized at $\boldsymbol{w}$ where the gradient is 0.

Bias term

This means, that all partial derivatives are 0

$$\Rightarrow \frac{\partial}{\partial w_1} \frac{1}{N_{TR}} \sum_{n=1}^{N_{TR}} \left| y_{TR}(n) - \sum_{m=1}^{M} \phi_m(\boldsymbol{x}_{TR}(n))\, w_{m+1} + w_1 \right|^2 = 0$$

Bias term

$$0 = \frac{\partial}{\partial w_1}\frac{1}{N_{TR}}\sum_{n=1}^{N_{TR}}\left(\left(y_{TR}(n) - \sum_{m=1}^{M}\phi_m(\boldsymbol{x}_{TR}(n))\,w_{m+1}\right)^2 + w_1^2 + 2\left(y_{TR}(n) - \sum_{m=1}^{M}\phi_m(\boldsymbol{x}_{TR}(n))\,w_{m+1}\right)w_1\right)$$

$$= \frac{1}{N_{TR}}\sum_{n=1}^{N_{TR}}\left(\frac{\partial}{\partial w_1}\left(y_{TR}(n) - \sum_{m=1}^{M}\phi_m(\boldsymbol{x}_{TR}(n))\,w_{m+1}\right)^2 + \frac{\partial}{\partial w_1}w_1^2 + 2\frac{\partial}{\partial w_1}\left(y_{TR}(n) - \sum_{m=1}^{M}\phi_m(\boldsymbol{x}_{TR}(n))\,w_{m+1}\right)w_1\right)$$

$$= \frac{1}{N_{TR}}\sum_{n=1}^{N_{TR}}\left(2w_1 + 2\left(y_{TR}(n) - \sum_{m=1}^{M}\phi_m(\boldsymbol{x}_{TR}(n))\,w_{m+1}\right)\right) = 2w_1 + \frac{2}{N_{TR}}\sum_{n=1}^{N_{TR}}\left(y_{TR}(n) - \sum_{m=1}^{M}\phi_m(\boldsymbol{x}_{TR}(n))\,w_{m+1}\right)$$

Bias term

$$\Rightarrow 2w_1 + \frac{2}{N_{TR}} \sum_{n=1}^{N_{TR}} \left(y_{TR}(n) - \sum_{m=1}^{M} \phi_m(\boldsymbol{x}_{TR}(n))\, w_{m+1} \right) = 0$$

$$\Rightarrow w_1 = \frac{1}{N_{TR}} \sum_{n=1}^{N_{TR}} \left(\sum_{m=1}^{M} \phi_m(\boldsymbol{x}_{TR}(n))\, w_{m+1} \right) - \frac{1}{N_{TR}} \sum_{n=1}^{N_{TR}} y_{TR}(n)$$

$$\Rightarrow w_1 = \frac{1}{N_{TR}} \sum_{n=1}^{N_{TR}} \left(\sum_{m=1}^{M} \phi_m(\boldsymbol{x}_{TR}(n))\, w_{m+1} \right) - \frac{1}{N_{TR}} \sum_{n=1}^{N_{TR}} y_{TR}(n)$$

“Bias” term w_1 compensates for difference between the average output and the average weighted sum of basis functions. Caution: this is **not the model estimation bias that we saw earlier.**

Maximum Likelihood Parameters

From MSE to LS – Reminder

Let D describe the data model.

Ideally, we'd like to have the solution to

$$\min_{\boldsymbol{W} \in R^{M+1}} E_{y,\boldsymbol{x}}\{|y - m(\boldsymbol{x}; \boldsymbol{w})|^2\}$$

Instead, we estimate m by $\widehat{m}$ and the MSE as

$$\min_{\boldsymbol{W} \in R^{M+1}} \|\boldsymbol{y} - \boldsymbol{H}^T \boldsymbol{w}\|_2^2$$

where $\boldsymbol{y} = [y_1, \dots, y_N]$ and $\boldsymbol{H} = [\boldsymbol{b}(\boldsymbol{x}_1), \boldsymbol{b}(\boldsymbol{x}_2), \dots, \boldsymbol{b}(\boldsymbol{x}_N)]$ depend on training data and $\widehat{m}$.

Least **Squares** (LS) estimates Mean **Squared** Error (MSE)

MSE (cont'd)

What about

$$\min_{\boldsymbol{W} \in R^{M+1}} E_{y,\boldsymbol{x}}\{|y - m(\boldsymbol{x}; \boldsymbol{w})| \quad \}$$

It could be estimated by

$$\min_{\boldsymbol{W} \in R^{M+1}} \|\boldsymbol{y} - \boldsymbol{H}^T \boldsymbol{w}\|_1$$

In general, why model dissimilarity as squared difference and not anything else?

Seems arbitrary…

Data Model Revisited

$$y = m(\boldsymbol{x}; \boldsymbol{w}) + \epsilon$$

Up until now, we had $E\{\epsilon\} = 0$ and $E\{\epsilon^2\} = V$.

This time, we also assume specific **error distribution**: $\epsilon \sim N(0, V)$

Why Gaussian?

Gaussian Distribution

$$X \sim N(\mu, \sigma^2)$$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

Central Limit Theorem (CLT): Under common conditions, the sum of many random variables will have an approximately normal distribution.

The normal distribution being the distribution with maximum entropy for a given mean and variance

That is, it makes the fewest assumptions on the RV.

Distribution of Output | Input

$$y = \boldsymbol{b}(\boldsymbol{x})^T \boldsymbol{w} + \epsilon$$

where $\epsilon \sim N(0, V)$

Then, for any given $\boldsymbol{x}$ and $\boldsymbol{w}$, y is distributed as:

$$N(\boldsymbol{b}(\boldsymbol{x})^T \boldsymbol{w},\ V)$$

Our goal remains to find $\boldsymbol{w}$

We will use **training data** (same as before), but also our **assumption on distribution** of $y|\boldsymbol{X}, \boldsymbol{w}$.

Likelihood

Likelihood is defined on a measurement, drawn from a distribution.

It measures "how likely this particular measurement is" in view of the distribution

For given model ($\hat{m}$ and $\boldsymbol{w}$) and any given input ($\boldsymbol{x}_i$), how likely is a particular output (y_i)?

$$f_*(y_i|\boldsymbol{x}_i, \boldsymbol{w})$$

Conditional PDF of y_i **given** $\boldsymbol{x}_i$ and $\boldsymbol{w}$ or the PDF of the conditioning defined RV $y_i|x_i, \boldsymbol{w}$

Maximum Likelihood Estimation

MLE Approach: Estimate $\widehat{\boldsymbol{w}}$ so that training outputs $\mathbf{y} = \left[y_1, y_2, \dots, y_{N_{TR}}\right]^{\mathrm{T}}$ appear to be the most likely ones, for the given training inputs $\boldsymbol{X} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{N_{TR}}\right]$.

We know that $(y_i|\boldsymbol{x}_i, \boldsymbol{w}) \sim N(\boldsymbol{w}^T \boldsymbol{b}(\boldsymbol{x}_i), V)$. That is:

$$f_*(y_i|\boldsymbol{x}_i, \boldsymbol{w}) = \frac{1}{\sqrt{2\pi V}} \exp\left(-\frac{1}{2} \frac{(y_i - \boldsymbol{b}(\boldsymbol{x}_i)^T \boldsymbol{w})^2}{V}\right)$$

But what about the PDF $\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w} \sim ?$

Maximum Likelihood Estimation

We assume that, given the inputs, the outputs are statistically independent. That is,

$$f_*(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w}) = f_*(\boldsymbol{y}|\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_{N_{TR}}\},\boldsymbol{w}) = \Pi_{i=1}^{N_{TR}} f_*(y_i|\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_{N_{TR}}\},w)$$

We also assume that y_i is independent of $\boldsymbol{x}_j$ for $i \neq j$. That is,

$$f_*(y_i|\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_{N_{TR}}\},\boldsymbol{w}) = f_*(y_i|\boldsymbol{x}_i,\boldsymbol{w})$$

Overall:

$$f_*(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w}) = \Pi_{i=1}^{N_{TR}} f_*(y_i|\boldsymbol{x}_i,\boldsymbol{w})$$

Maximum Likelihood Estimation

$$f_*(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w}) = \Pi_{i=1}^{N_{TR}} \frac{1}{\sqrt{2\pi V}} \exp\left(-\frac{1}{2}\,\frac{(y_i - \boldsymbol{b}(\boldsymbol{x}_i)^T\boldsymbol{w})^2}{V}\right) = \frac{1}{\left(\sqrt{2\pi V}\right)^{N_{TR}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_{TR}} \frac{(y_i - \boldsymbol{b}(\boldsymbol{x}_i)^T\boldsymbol{w})^2}{V}\right)$$

$$f_*(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w}) = \frac{1}{\left(\sqrt{2\pi V}\right)^{N_{TR}}} \exp\left(-\frac{1}{2V}\ \|\boldsymbol{y} - \boldsymbol{H}^T\boldsymbol{w}\|_2^2\right) = N(\boldsymbol{H}^T\boldsymbol{w}, V\,\boldsymbol{I}_N)$$

Note: $\boldsymbol{x} \sim N(\boldsymbol{m},\boldsymbol{C}) \Leftrightarrow f(\boldsymbol{x}) = \ det(2\,\pi\,\boldsymbol{C})^{-\frac{1}{2}} exp\left(-\frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{m})\right)$

Maximum Likelihood Estimation

MLE: Estimate $\widehat{\boldsymbol{w}}$ so that training outputs $\mathbf{y} = \left[y_1, y_2, \dots, y_{N_{TR}}\right]^{\mathrm{T}}$ appear to be the most likely ones,

given training inputs $\boldsymbol{X} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{N_{TR}}\right]$.

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in R^{M+1}}{\arg\max}\, f_{\mathrm{PDF}}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$$

So, we want to solve:

$$\max_{w \in R^{M+1}} \frac{1}{\left(\sqrt{2\pi V}\right)^{N_{TR}}} \exp\left(-\frac{1}{2V} \|\boldsymbol{y} - \boldsymbol{H}^T \boldsymbol{w}\|_2^2\right)$$

Note on Optimization

- **Two metrics are minimized/maximized by the same argument**, if one is a monotonically increasing function of the other.
- Examples of monotonically increasing functions: $h(\boldsymbol{x}) = 10C(\boldsymbol{x})$, $h(\boldsymbol{x}) = C(\boldsymbol{x})^2$, $h(\boldsymbol{x}) = \sqrt{C(\boldsymbol{x})}$, $h(\boldsymbol{x}) = \log\big(C(\boldsymbol{x})\big)$, $h(\boldsymbol{x}) = \exp\big(C(\boldsymbol{x})\big)$.

- In cases like the above, $\min_{\boldsymbol{w}} C(x)$and $\min_{\boldsymbol{w}} h(x)$ are called equivalent problems.
- When $h(\boldsymbol{x})$ is a decreasing function of $C(\boldsymbol{x})$, then the arguments that minimize $h(\boldsymbol{x})$ also maximize $C(\boldsymbol{x})$ (and vice versa).

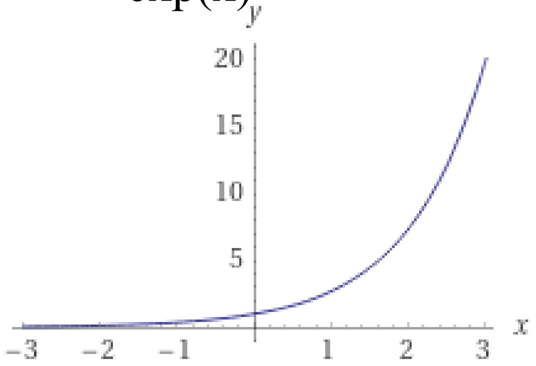
- Examples of decreasing functions: $h(\boldsymbol{x}) = -10C(\boldsymbol{x})$, $h(\boldsymbol{x}) = C(\boldsymbol{x})^{-2}$, etc.

Maximum Likelihood Estimation

Accordingly, the ML problem

$$\max_{\boldsymbol{w}\in R^{M+1}} \frac{1}{\left(\sqrt{2\pi V}\right)^{N_{TR}}} \exp\left(-\frac{1}{2V}\ \|\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w}\|_2^2\right)$$

exp(x)



is equivalent to the simpler:

$$\min_{\boldsymbol{w}\in R^{M+1}} \|\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w}\|_2^2$$

Which means:

$$\text{MLE} \overset{\text{Gaussian Error}}{\equiv} \text{LS}$$

MLE

Which means:

$$\text{MLE} \overset{\text{Gaussian Error}}{\equiv} \text{LS}$$

Which means that MSE-LS has likelihood optimality.

The MSE-LS estimate is the most likely one.

For the reasonable case of Normal noise, MSE-LS is **maximum-likelihood optimal**.

What about other distribution assumptions on ϵ? What if Gaussian but correlated across training data?

Shortcomings of MLE

- Same as MSE – LS
- To control overfitting, you must increase N or decrease M

Bayesian Regression

Bayes Rule

For events A and B:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Offers a way to swap conditioning.

For discrete/continuous random variables $\underline{X} \in S_X, \underline{Z} \in S_Z$, simplify PMF/PDF notation as

$$p_{\underline{X}}(X) \text{ or } f_{\underline{X}}(X) \rightarrow f_*(X)$$

$$p_{\underline{X},\underline{Z}}(X,Z) \text{ or } f_{\underline{X},\underline{Z}}(X,Z) \rightarrow f_*(X,Z)$$

$$p_{\underline{X}|\underline{Z}}(X,Z) \text{ or } f_{\underline{X}|\underline{Z}}(X,Z) \rightarrow f_*(X|Z)$$

Bayes Rule becomes:

$$f_*(X|Z) = \frac{f_*(Z|X)f_*(X)}{f_*(Z)} \; \forall (X,Z) \in S_X \times S_Z$$

Random Parameters

Until now parameter vector $\boldsymbol{w}$ was seen as deterministic optimization argument.

The best configuration of $\boldsymbol{w}_{\text{best}}$ is unknown. Therefore, we can see it as a random vector.

Thus, it must have some density function, $f_*(\boldsymbol{w})$.

If f_* was known, how would you choose your parameters?

Choose $\boldsymbol{w}$ that exhibits the highest probability to be the best.

That is, choose the argument that solves:

$$\max_{\boldsymbol{w}} f_*(\boldsymbol{w})$$

Prior Distribution

Can we assume that f_* (the PDF of the best parameter configuration) is known?

Probably not really.

But we can assume that we have a good initial (=we call it "prior") guess of it.

From now on, f_* denotes our prior guess on the distribution of the best $\boldsymbol{w}$.

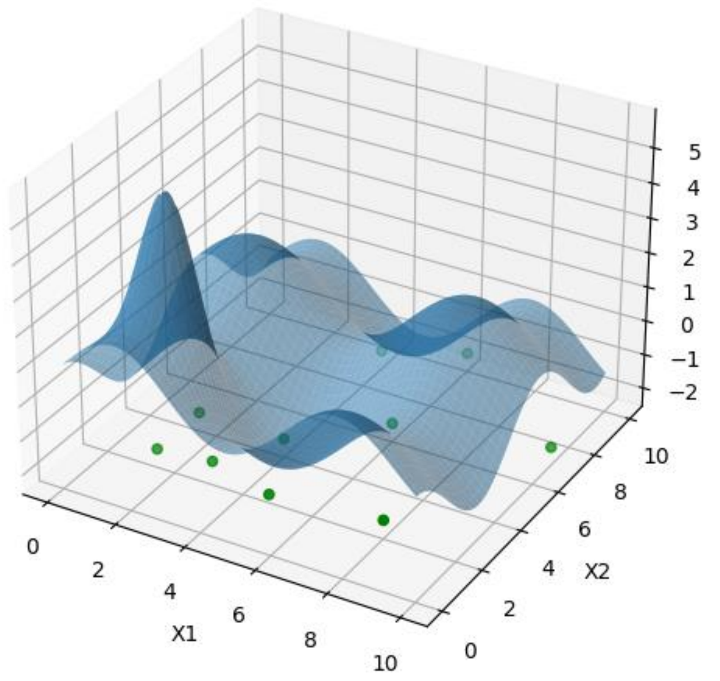
For example, we can assume

$$f_*(\boldsymbol{w}) = N(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0)$$

for some $\boldsymbol{m}_0, \boldsymbol{\Sigma}_0$, since Gaussian makes the fewest assumptions.

Even simpler, we can assume zero-mean isotropic Gaussian $f_*(\boldsymbol{w}) = N(\boldsymbol{0}, a\boldsymbol{I})$

Prior Distribution (cont'd)



Example: Considering GRBF model, if we have a guess or expert knowledge that in our data model output y takes higher values when the input is around $[2,2]^T$, then I can assume that the parameters weighting the GRBF-centers that are near $[2,2]^T$ would take higher values with higher probability.

Prior Distribution (cont'd)

So f_* denotes our prior guess on the distribution of the best $\boldsymbol{w}$.

Again, we can choose our parameters to be the solution to $\max_{\boldsymbol{w}} f_*(\boldsymbol{w})$

If $f_*(\boldsymbol{w}) = N(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0)$, $\widehat{\boldsymbol{w}} = \boldsymbol{m}_0$, since Gaussian bell attains max value at the mean.

But maybe I can use some training data to help us improve my best-parameter distribution guess.

Let $(\boldsymbol{y}, \boldsymbol{X})$ be the training dataset ($y_i = [\boldsymbol{y}]_i$ is the output corresponding to input $[\boldsymbol{X}]_{:,i}$).

Our updated guess of the best-parameter distribution, after training data $(\boldsymbol{y}, \boldsymbol{X})$ have been studied, we call it "posterior" or "a posteriori" distribution.

- **Prior** guess on best-parameter distribution: $f_*(\boldsymbol{w})$
- **Posterior** guess on best-parameter distribution: $f_*(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})$

Posterior Distribution

But maybe I can use some training data to help us improve my best-parameter distribution guess.

Let $(\boldsymbol{y}, \boldsymbol{X})$ be the training dataset ($y_i = [\boldsymbol{y}]_i$ is the output corresponding to input $[\boldsymbol{X}]_{:,i}$).

Our updated guess of the best-parameter distribution, after training data $(\boldsymbol{y}, \boldsymbol{X})$ have been studied, we call it "posterior" or "a posteriori" distribution.

- **Prior** guess on best-parameter distribution: $f_*(\boldsymbol{w})$
- **Posterior** guess on best-parameter distribution: $f_*(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})$

Maximum A Posteriori Probability

When we had only the prior distribution, we designed our parameters by solving:

$$\max_{\boldsymbol{w}} f_*(\boldsymbol{w})$$

If we use the data in $(\boldsymbol{y}, \boldsymbol{X})$ a smart way and derive the posterior distribution/probability $f_*(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})$, then we'll design our parameters by solving:

$$\max_{\boldsymbol{w}} f_*(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})$$

This is called the **Maximum A Posteriori Probability** approach.

Two steps follow:

Step 1: $f_*(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})$

Step 2: Solve $max_{\boldsymbol{w}}\, f_*(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X})$

Posterior Distribution

We are looking for posterior distribution $f_*(\boldsymbol{w}\,|\boldsymbol{y}, \boldsymbol{X})$.

This means that $\boldsymbol{w}$ is the variable and that $\boldsymbol{y}, \boldsymbol{X}$ are given (conditioned over) and fixed.

Therefore, terms like $f_*(\boldsymbol{X})$ and $f_*(\boldsymbol{y}, \boldsymbol{X})$ are just positive constants with respect to $\boldsymbol{w}$.

$$f_*(\boldsymbol{w}\,|\boldsymbol{y}, \boldsymbol{X}) = \frac{f_*(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{w}) f_*(\boldsymbol{w})}{f_*(\boldsymbol{y}, \boldsymbol{X})} \propto f_*(\boldsymbol{y}, \boldsymbol{X}|\boldsymbol{w}) f_*(\boldsymbol{w})$$

$$\propto f_*(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) f_*(\boldsymbol{X}|\boldsymbol{w}) f_*(\boldsymbol{w})$$

$$\propto f_*(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) f_*(\boldsymbol{X}) f_*(\boldsymbol{w})$$

$$\propto f_*(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) f_*(\boldsymbol{w})$$

Thus, $f_*(\boldsymbol{w}\,|\boldsymbol{y}, \boldsymbol{X}) = f_*(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})\, f_*(\boldsymbol{w}) \cdot \text{pos\_const}$

Posterior Distribution

$$f_*(\boldsymbol{w}\,|\boldsymbol{y},\boldsymbol{X}) = f_*(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})\, f_*(\boldsymbol{w}) \cdot \text{pos\_const}$$

During our MLE studies we we found:

$$f_*(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w}) = N(\boldsymbol{H}^T\boldsymbol{w}, V\,\boldsymbol{I}_N) = \text{pos\_const} \cdot \exp\left(-\frac{1}{2}\,(\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w})^T(V\,\boldsymbol{I}_N)^{-1}(\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w})\right)$$

About our prior, we assume Gaussian (fewest assumptions) with specific mean and covariance:

$$f_*(\boldsymbol{w}) = N(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0) = \text{pos\_const} \cdot exp\left(-\frac{1}{2}\,(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)\right)$$

Posterior Distribution (cont'd)

Thus,

$$f_*(\boldsymbol{w}\,|\boldsymbol{y},\boldsymbol{X}) = f_*(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})\, f_*(\boldsymbol{w}) \cdot \text{pos\_const}$$

$$= \text{pos\_const} \cdot exp\left(\left(-\frac{1}{2}\,(\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w})^T(V\,\boldsymbol{I}_N)^{-1}(\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w})\right)+\left(-\frac{1}{2}\,(\boldsymbol{w}-\boldsymbol{m}_w)^T(\boldsymbol{S}_w)^{-1}(\boldsymbol{w}-\boldsymbol{m}_w)\right)\right)$$

$$= \text{pos\_const} \cdot exp\left(-\frac{1}{2}\left((\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w})^T(V\,\boldsymbol{I}_N)^{-1}(\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w})+(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)\right)\right)$$

Posterior Distribution (cont'd)

Make sure you know how to derive:

$$(\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w})^T(V\,\boldsymbol{I}_N)^{-1}(\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w})+(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)$$

$$=(\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w})+(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)$$

$$=\boldsymbol{y}^T\boldsymbol{V}^{-1}\boldsymbol{y}+\boldsymbol{w}^T\boldsymbol{H}\ \ \boldsymbol{V}^{-1}\boldsymbol{H}^T\boldsymbol{w}-2\boldsymbol{w}^T(\boldsymbol{H}\boldsymbol{V}^{-1}\boldsymbol{y})+\boldsymbol{w}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{w}+\boldsymbol{m}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0-2\boldsymbol{w}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0$$

$$=\boldsymbol{y}^T\boldsymbol{V}^{-1}\boldsymbol{y}+\boldsymbol{m}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0+\boldsymbol{w}^T(\boldsymbol{H}\boldsymbol{V}^{-1}\boldsymbol{H}^T+\boldsymbol{\Sigma}_0^{-1})\boldsymbol{w}-2\boldsymbol{w}^T(\boldsymbol{H}\boldsymbol{V}^{-1}\boldsymbol{y}+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0)$$

$$=\boldsymbol{y}^T\boldsymbol{V}^{-1}\boldsymbol{y}+\boldsymbol{m}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0+\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{z}$$

$$=\boldsymbol{y}^T\boldsymbol{V}^{-1}\boldsymbol{y}+\boldsymbol{m}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0+\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\,\boldsymbol{z}$$

$$=\boldsymbol{y}^T\boldsymbol{V}^{-1}\boldsymbol{y}+\boldsymbol{m}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0+\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{m}+\boldsymbol{m}^T\boldsymbol{m}\ -\boldsymbol{m}^T\boldsymbol{m}$$

$$=\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w}-2\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{m}+\boldsymbol{m}^T\boldsymbol{m}+\text{pos\_const}$$

$$=(\boldsymbol{w}-\boldsymbol{m})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{m})+\text{pos\_const}$$

Posterior Distribution (cont'd)

$$f_*(\boldsymbol{w}\,|\boldsymbol{y},\boldsymbol{X}) = \beta \cdot exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{m})\right) = \boldsymbol{N}(\boldsymbol{m},\boldsymbol{\Sigma})$$

$\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X}$ is multivariate Gaussian with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma} = \left(\frac{1}{V}\boldsymbol{H}\boldsymbol{H}^T + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

$$\boldsymbol{m} = \boldsymbol{\Sigma}\left(\frac{1}{V}\boldsymbol{H}\quad \boldsymbol{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0\right) = (\boldsymbol{H}\boldsymbol{H}^T + V\boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{H}\boldsymbol{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0\, V)$$

Maximize Posterior Density

$$f_*(\boldsymbol{w}\,|\boldsymbol{y},\boldsymbol{X}) = \beta \cdot exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{m})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{m})\right) \;= \boldsymbol{N}(\boldsymbol{m},\boldsymbol{\Sigma})$$

Gaussian PDF is maximized at its mean.

That is, the MAP parameter vector is

$$\widehat{\boldsymbol{w}} = (\boldsymbol{H}\boldsymbol{H}^T + V\boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{H}\boldsymbol{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0\, V)$$

Read is data, **blue** is prior distribution.

Alternative Derivation

We showed that $\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X}$ follows $\boldsymbol{N}(\boldsymbol{m},\boldsymbol{\Sigma})$. We recognized that MAP parameters should be $\boldsymbol{m}$.

Here is an alternative derivation, without finding first the distribution of $\boldsymbol{w}|\boldsymbol{y},\boldsymbol{X}$.

$$\max_{\boldsymbol{w}} f_*(\boldsymbol{w}\,|\boldsymbol{y},\boldsymbol{X})$$

$$\equiv \max_{\boldsymbol{w}} f_*(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})f_*(\boldsymbol{w})\cdot \text{pos\_onst.}$$

$$\equiv \max_{\boldsymbol{w}} f_*(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})f_*(\boldsymbol{w})$$

$$\equiv \max_{\boldsymbol{w}} \ln\big(f_*(\boldsymbol{y},\boldsymbol{X}|\boldsymbol{w})\big)+\ln\big(f_*(\boldsymbol{w})\big)$$

Posterior Density Maximization

$$\max_{\boldsymbol{w}} f_*(\boldsymbol{w}\,|\boldsymbol{y},\boldsymbol{X})$$

$$\equiv \max_{\boldsymbol{w}} \ln\left(\frac{1}{\left(\sqrt{2\pi V}\right)^N}\exp\left(-\frac{1}{2V}\;\|\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w}\|_2^2\right)\right) + \ln\left(\frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_0\,|}}\exp\left(-\frac{1}{2}\;(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)\right)\right)$$

$$\equiv \max_{\boldsymbol{w}} \ln\left(\frac{1}{\left(\sqrt{2\pi V}\right)^N}\right) - \frac{1}{2V}\;\|\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w}\|_2^2 + \ln\left(\frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_0\,|}}\right) - \frac{1}{2}\;(\boldsymbol{w}-\boldsymbol{m}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)$$

$$\max_{\boldsymbol{w}} f_*(\boldsymbol{w}\,|\boldsymbol{y},\boldsymbol{X}) \;\equiv \min_{\boldsymbol{w}}\;\|\boldsymbol{y}-\boldsymbol{H}^T\boldsymbol{w}\|_2^2 \;+(\boldsymbol{w}-\boldsymbol{m}_0)^T(V\boldsymbol{\Sigma}_0^{-1})(\boldsymbol{w}-\boldsymbol{m}_0)$$

Posterior Density Maximization (cont'd)

$$\max_{\boldsymbol{w}} \ \|\boldsymbol{y} - \boldsymbol{H}^T\boldsymbol{w}\|_2^2 \ + (\boldsymbol{w} - \boldsymbol{m}_0)^T(V\boldsymbol{\Sigma}_0^{-1})(\boldsymbol{w} - \boldsymbol{m}_0)$$

Convex objective function. Apply stationarity condition:

$$C(\boldsymbol{w}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{H}\boldsymbol{y} + \boldsymbol{w}^T(V\boldsymbol{\Sigma}_0^{-1})\boldsymbol{w} + \boldsymbol{m}_0{}^T(V\boldsymbol{\Sigma}_0^{-1})\boldsymbol{m}_0 - 2\boldsymbol{w}^T(V\boldsymbol{\Sigma}_0^{-1})\boldsymbol{m}_0$$

$$\boldsymbol{g}(\boldsymbol{w}) = 2\ \boldsymbol{H}\boldsymbol{H}^T\boldsymbol{w} - 2\boldsymbol{H}\boldsymbol{y} + 2(V\boldsymbol{\Sigma}_0^{-1})\boldsymbol{w} - 2\ (V\boldsymbol{\Sigma}_0^{-1})\boldsymbol{m}_0$$

$$\boldsymbol{g}(\widehat{\boldsymbol{w}}) = \boldsymbol{0} \Leftrightarrow$$

$$\widehat{\boldsymbol{w}} = (\boldsymbol{H}\boldsymbol{H}^T + V\boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{H}\boldsymbol{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0\ V)$$

MAP – Special Cases

MAP solution:

$$\hat{\boldsymbol{w}} = (\boldsymbol{H}\boldsymbol{H}^T + V\boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{H}\boldsymbol{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0 V)$$

No data? Boils down to prior distribution maximization.

$$\boldsymbol{H} = 0 \Rightarrow \hat{\boldsymbol{w}} = (\boldsymbol{0} + V\boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{0} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{m}_0 V) = \boldsymbol{m}_0$$

No prior distribution assumption? Boils done to MLE/MSE/LS.

Equivalent to considering $\boldsymbol{\Sigma}_0 = a\boldsymbol{I}$ with $a \to \infty$.

$$a \to \infty \Rightarrow \boldsymbol{\Sigma}_0^{-1} = \boldsymbol{0} \Rightarrow \hat{\boldsymbol{w}} = (\boldsymbol{H}\boldsymbol{H}^T + \boldsymbol{0})^{-1}(\boldsymbol{H}\boldsymbol{y} + \boldsymbol{0}) = (\boldsymbol{H}\boldsymbol{H}^T)^{-1}\boldsymbol{H}\boldsymbol{y}$$

MLE vs MAP

- Same as comparison between LS and Regularized-LS.
- MLE relies solely on data and it needs many, otherwise might overfit.
- MAP with Gaussian error and prior is LS + regularization.
- MAP relies also on model priors; less prone to overfit; could work with less data.
- Prior with very high variance: MAP tends to MLE/MSE/LS
- Limited/no data? MAP tends to prior mean.