

Exercise 5

Machine Learning I

4A-1.

We know:

$$p(\theta) = \frac{1}{\Delta_{prior}}$$
$$p(\theta_{MAP}|D) \approx \frac{1}{\Delta_{posterior}}$$

Using Bayes' theorem and rearranging terms we get:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$
$$(*) \approx p(D) = p(D|\theta) \frac{\Delta_{posterior}}{\Delta_{prior}}$$

As we saw in the lecture slides, the evidence approximation is useful to make inferences about model complexity and fit.

Let us see what happens in (*) if we vary the input D :

As the likelihood $p(D|\theta)$ is strongly related to the fit, $p(D)$ acquires information about the model fit. As fit is related to model complexity, a better fit usually entails a more complex model as well. Additionally, because of $\frac{\Delta_{posterior}}{\Delta_{prior}}$ our $p(D)$ also receives information about model complexity and plausibility.

If $\Delta_{posterior}$ is too high compared to our belief Δ_{prior} , then $p(D)$ goes down. Think about what this means. A very high posterior around θ_{MAP} leads to a point estimate not unlike the maximum likelihood method. This leads to little bias in the input set \mathbf{x}_i . As we know from the bias/variance decomposition, this may increase variance for different datasets \mathbf{x}_j . Which means we have learned too much from our data.

Because $p(D)$ goes down in this case, it helps us to combat overfitting.

On the other hand, if our fit is good ($p(D|\theta)$ high), yet we have a general enough model ($p(\theta_{MAP}|D)$ low), we receive a good value for $p(D)$.

Caution: As we receive more data points, $p(\theta_{MAP}|D)$ usually increases because the variance in our posterior decreases. This makes us more liable to overfit, as additional data also increases the configuration space.

The optimal value for $p(D)$ is thereby a balance of model complexity, belief ($\frac{\Delta_{posterior}}{\Delta_{prior}}$) and fit ($p(D|\theta)$).

4A-2.

Prerequisites

Let \mathbf{A} be invertible, \mathbf{B}, \mathbf{C} arbitrary and \mathbf{I} be the identity matrix.

We then have:

$$(*) \quad (\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - (\mathbf{I} + \mathbf{A}^{-1}\mathbf{BCD})^{-1}\mathbf{A}^{-1}\mathbf{BCDA}^{-1}$$

The proof of above can be seen here:

<http://www0.cs.ucl.ac.uk/staff/gridgway/mil/mil.pdf>

Also, as seen in the last lecture slides, the posterior of Bayesian linear regression with normal prior is:

$$p(\mathbf{w}|\mathbf{t}, \Sigma_N, \tau_e) = N(\mathbf{w}; \boldsymbol{\mu}_N, \Sigma_N)$$

$$\boldsymbol{\mu}_N = \tau_e \Sigma_N \Phi^T \mathbf{t}$$

$$\Sigma_N = (\Sigma_0^{-1} + \tau_e \Phi^T \Phi)^{-1}$$

Let us have K old and M “new” datapoints.

Because the predictive distribution is dependent on the posterior, it makes sense to calculate the new posterior first.

As seen in the task, we have:

$$p(\mathbf{w}|\mathbf{t}_M, \mathbf{t}_K) \propto p(\mathbf{w}|\mathbf{t}_K)p(\mathbf{t}_M|\mathbf{w})$$

This leads to:

$$p(\mathbf{w}|\mathbf{t}_K)p(\mathbf{t}_M|\mathbf{w}) \propto e^{-\frac{1}{2}[(\mathbf{w}-\boldsymbol{\mu}_K)^T \Sigma_K^{-1}(\mathbf{w}-\boldsymbol{\mu}_K) + (\mathbf{t}_M - \Phi_M \mathbf{w})^T \tau_e \mathbf{I}(\mathbf{t}_M - \Phi_M \mathbf{w})]}$$

Note: We have $\tau_e \mathbf{I}$ because the predictions are independent from each other. Focusing on the exponent:

$$\begin{aligned}
(\mathbf{w} - \boldsymbol{\mu}_K)^T \boldsymbol{\Sigma}_K^{-1} (\mathbf{w} - \boldsymbol{\mu}_K) + (\mathbf{t}_M - \boldsymbol{\Phi}_M \mathbf{w})^T \tau_e \mathbf{I} (\mathbf{t}_M - \boldsymbol{\Phi}_M \mathbf{w}) &= \mathbf{w}^T \underbrace{(\boldsymbol{\Sigma}_K^{-1} + \tau_e \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M)}_{\boldsymbol{\Sigma}_{K+M}^{-1}} \mathbf{w} - 2\mathbf{w}^T [\boldsymbol{\Sigma}_K^{-1} \boldsymbol{\mu}_K + \tau_e \boldsymbol{\Phi}_M^T \mathbf{t}_M] \\
&\quad + \boldsymbol{\mu}_K^T \boldsymbol{\Sigma}_K^{-1} \boldsymbol{\mu}_K + \tau_e \mathbf{t}_M^T \mathbf{t}_M \\
&= \mathbf{w}^T \boldsymbol{\Sigma}_{K+M}^{-1} \mathbf{w} \\
&\quad - 2\mathbf{w}^T \underbrace{\boldsymbol{\Sigma}_{K+M}^{-1} \boldsymbol{\Sigma}_{K+M}}_{=1} [\boldsymbol{\Sigma}_K^{-1} \boldsymbol{\mu}_K + \tau_e \boldsymbol{\Phi}_M^T \mathbf{t}_M] \\
&\quad + c
\end{aligned}$$

This allows us to complete the square above and we get the updated posterior:

$$p(\mathbf{w} | \mathbf{t}_M, \mathbf{t}_K) = N(\boldsymbol{\mu}_{M+K}, \boldsymbol{\Sigma}_{K+M})$$

with

$$\begin{aligned}
\boldsymbol{\Sigma}_{K+M} &= (\boldsymbol{\Sigma}_K^{-1} + \tau_e \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M)^{-1} \\
\boldsymbol{\mu}_{M+K} &= \boldsymbol{\Sigma}_{K+M} [\boldsymbol{\Sigma}_K^{-1} \boldsymbol{\mu}_K + \tau_e \boldsymbol{\Phi}_M^T \mathbf{t}_M]
\end{aligned}$$

Note: When $\boldsymbol{\mu}_K = \mathbf{0}$ we get the same result as in the slides for the prior $p(\mathbf{w} | \boldsymbol{\Sigma}_0)$.

Now we use (*) for the covariance:

$$(\boldsymbol{\Sigma}_K^{-1} + \tau_e \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M)^{-1} = \boldsymbol{\Sigma}_K - \tau_e (\mathbf{I} + \tau_e \boldsymbol{\Sigma}_K \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M)^{-1} \boldsymbol{\Sigma}_K \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M \boldsymbol{\Sigma}_K$$

As seen in the slides (or the task), we try to show:

$$\sigma_{K+M}(x^*) \leq \sigma_K(x^*)$$

where

$$\sigma_n = \frac{1}{r_e} + \boldsymbol{\phi}(x^*)^T \boldsymbol{\Sigma}_N \boldsymbol{\phi}(x^*)$$

Now we just plug our new covariance matrices into σ_n :

$$\begin{aligned}
\sigma_{K+M}(x^*) &= \frac{1}{r_e} \boldsymbol{\phi}(x^*)^T (\boldsymbol{\Sigma}_K^{-1} + \tau_e \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M)^{-1} \boldsymbol{\phi}(x^*) \\
&= \frac{1}{r_e} \boldsymbol{\phi}(x^*)^T [\boldsymbol{\Sigma}_K - \tau_e (\mathbf{I} + \tau_e \boldsymbol{\Sigma}_K \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M)^{-1} \boldsymbol{\Sigma}_K \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M \boldsymbol{\Sigma}_K] \boldsymbol{\phi}(x^*) \\
&= \underbrace{\frac{1}{r_e} \boldsymbol{\phi}(x^*)^T \boldsymbol{\Sigma}_K \boldsymbol{\phi}(x^*)}_{\sigma_K(x^*)} - \frac{1}{r_e} \boldsymbol{\phi}(x^*)^T [\tau_e (\mathbf{I} + \tau_e \boldsymbol{\Sigma}_K \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M)^{-1} \boldsymbol{\Sigma}_K \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M \boldsymbol{\Sigma}_K] \boldsymbol{\phi}(x^*) \\
&\leq \sigma_K(x^*).
\end{aligned}$$

The last equality holds because

$$\tau_e (\mathbf{I} + \tau_e \boldsymbol{\Sigma}_K \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M)^{-1} \boldsymbol{\Sigma}_K \boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M \boldsymbol{\Sigma}_K$$

is at least positive semidefinite (for why, look up the properties of positive semidefinite matrices, especially in regard to $\boldsymbol{\Sigma}_K$ and $\boldsymbol{\Phi}_M^T \boldsymbol{\Phi}_M$).

4A-3.

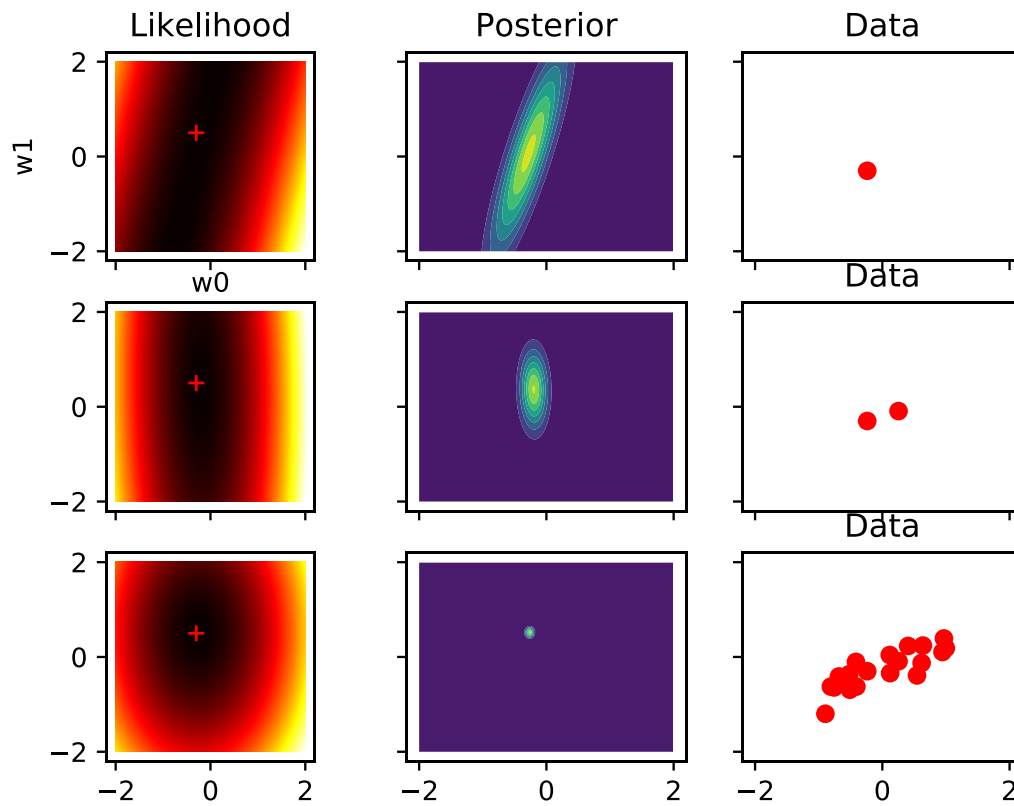


Figure 1 Replication of Bishop's Figure 3.7 in Python