Machine Learning
SS 2020
Exercise sheet 2

Solution by
Lorenzo Minecci, Daniel Strenger
11939539, 01531211

May 11, 2020

Exercise 1. (Online Bayesian Linear Regression)
Solution. 1.

$$p(t|\boldsymbol{x},\boldsymbol{w}) = \prod_{n=1}^{N} \sqrt{\frac{\beta}{2\pi}} \exp{(-\beta\frac{(\boldsymbol{t}-\boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2}{2})}$$

$$\log(p(\boldsymbol{t}|\boldsymbol{x},\boldsymbol{w})) = \sum_{n=1}^{N}\frac{1}{2}(\ln\beta - \ln 2\pi) + [-\frac{\beta}{2}(\boldsymbol{t}-\boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2] = \frac{N}{2}(\ln\beta-\ln 2\pi) - \beta\underbrace{\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{t}-\boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2}_{E_D} =$$

$$\frac{N}{2}(\ln\beta-\ln 2\pi) - \beta E_D$$

Solution. 2.
It's a multivariate Gaussian, and again we are going to compute the logarithm. So:

$$\begin{aligned}\ln(p(\boldsymbol{w})) = \ln(\mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\alpha^{-1}I)) = \ln\left(\frac{1}{(2\pi)^{\frac{M+1}{2}}|\alpha^{-1}I|}\exp\left((-\frac{1}{2}\boldsymbol{w}^T(\alpha^{-1}I)\boldsymbol{w})\right)\right) &= \\ -\frac{M+1}{2}\ln(2\pi) - \frac{1}{2}\ln(|\alpha^{-1}I|) - \left(\frac{1}{2}\boldsymbol{w}^T(\alpha^{-1}I)\boldsymbol{w}\right) &= \\ -\frac{M+1}{2}\ln(2\pi) + \frac{M+1}{2}\ln(\alpha) - \frac{1}{2}\sum_{i=1}^{M+1}\frac{w_i}{\alpha}\end{aligned}$$

In this setting we notice that only the third term is dependent on w, while the first two terms as constants. We'll use this consideration for the next exercise.

Solution. 3.
Bayes theorem states that:

$$p(\boldsymbol{w}|\boldsymbol{t}) = \frac{p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{t})}$$

We assume that $p(\boldsymbol{t})$ is not depending on w, so also its logarithm will be independent.
Let's compute the logarithm of the posterior simply as:

$$\begin{aligned}\ln(p(\boldsymbol{t}|\boldsymbol{w})) + \ln(p(\boldsymbol{w})) + \mathit{cost} = \underbrace{\frac{N}{2}(\ln\beta - \ln(2\pi))}_{\text{cost}} - \frac{\beta}{2}\sum_{n=1}^{N}(\boldsymbol{t} - \boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2 + \\ + \left[\underbrace{-\frac{D}{2}\ln(2\pi) - \frac{D}{2}\ln(\alpha)}_{\text{cost}} + \left(-\frac{1}{2}\boldsymbol{w}^T(\alpha^{-1}I)^{-1}\boldsymbol{w}\right)\right] =\end{aligned}$$

concluding:

$$\ln(p(\boldsymbol{w}|\boldsymbol{t})) = -\frac{\beta}{2}\sum_{n=1}^{N}(\boldsymbol{t} - \boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \mathit{cost}$$

where we grouped in *cost* all the terms constant with respect to w. The alpha becomes a scalar multiplying the vector product. The maximization of this distribution w.r.t. $\boldsymbol{w}$ is equivalent to the minimization of the sum-of-squares approach with addition of a quadratic regularization term $\lambda = \frac{\alpha}{\beta}$.

Solution. 4.
Let's now maximise the log-posterior w.r.t. w We can express the vectors $\phi(x_i)$ as rows of the design matrix Φ

$$\begin{aligned}\frac{\partial}{\partial w_i}\left[\frac{\beta}{2}\sum_{n=1}^{N}(\boldsymbol{t} - \boldsymbol{w}^T\boldsymbol{\phi}(x_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \mathit{cost}\right] &= \frac{\beta}{2}(2\Phi_i^T\Phi\boldsymbol{w} - 2\Phi_i^T\boldsymbol{t}) + \frac{\alpha}{2}2\boldsymbol{w} = \\ &= \left(\beta\Phi^T\Phi + \alpha I\right)\boldsymbol{w} - \beta\Phi_i^T\boldsymbol{t} \overset{!}{=} 0\end{aligned}$$

$$\Rightarrow \boldsymbol{w}_{map} = (\Phi^T\Phi + \frac{\alpha}{\beta}I)^{-1}\Phi\boldsymbol{t}$$

That we can see it's equal to $\boldsymbol{m} = \beta\boldsymbol{S\Phi^T t}$, being $\boldsymbol{S} = \left(\alpha\boldsymbol{I} + \beta\boldsymbol{\Phi^T\Phi}\right)^{-1}$.
We took β out of $\boldsymbol{S}$, which is finally a scalar multiplier in $\boldsymbol{m}$.

Exercise 2. (Logistic Regression)
Solution. 1. We have $p(t=1|\tilde{\boldsymbol{x}}) = \sigma(\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}) = \sigma(1\cdot\boldsymbol{w}^t\tilde{\boldsymbol{x}})$ and

$$\begin{aligned}p(t=-1|\tilde{\boldsymbol{x}}) = 1 - \sigma(\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}) = 1 - \frac{1}{1+e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}} &= \frac{1+e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}}{1+e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}} - \frac{1}{1+e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}} \\ &= \frac{e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}}{1+e^{-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}} = \frac{1}{e^{\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}}}+1} = \sigma(-\tilde{\boldsymbol{w}}^t\tilde{\boldsymbol{x}})\end{aligned}$$

2.

$$
\begin{aligned}
-\log(p(\tilde{\boldsymbol{w}}|\boldsymbol{t},\boldsymbol{x})) = -\log(p(\boldsymbol{w})) - \sum_{n=1}^{N} \log(\sigma(t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}})) = \\
-\log(p(\boldsymbol{w})) - \sum_{n=1}^{N} \log(\frac{1}{1+e^{-t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}}}) \\
= -\log(p(\boldsymbol{w})) + \sum_{n=1}^{N} \log(1+e^{-t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}}) = \\
-\log\left(\frac{1}{(2\pi S^2)^{N/2}} \exp\left(-\frac{1}{2S^2}\sum_{n=1}^{N} w_n^2\right)\right) + \sum_{n=1}^{N} \log(1+e^{-t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}}) \\
= -\log\left(\frac{1}{(2\pi S^2)^{N/2}}\right) + \frac{1}{2S^2}\sum_{n=1}^{N} w_n^2 + \sum_{n=1}^{N} \log(1+e^{-t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}})
\end{aligned}
$$

3.

$$
\begin{aligned}
&\frac{\partial}{\partial b} E(\tilde{\boldsymbol{w}}) = \sum_{n=1}^{N} \frac{\partial}{\partial b} \log(1+e^{-t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}}) = \sum_{n=1}^{N} \frac{-t_n e^{-t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}}}{1+e^{-t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}}}, \\
&\frac{\partial}{\partial w_i} E(\tilde{\boldsymbol{w}}) = \frac{1}{2S^2} \sum_{n=1}^{N} \frac{\partial}{\partial w_i} w_n^2 + \sum_{n=1}^{N} \frac{\partial}{\partial w_i} \log(1+e^{-t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}}) \\
&\qquad = \frac{w_i}{S^2} + \sum_{n=1}^{N} \frac{-t_n \tilde{x}_n^i e^{-t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}}}{1+e^{-t_n \tilde{\boldsymbol{w}}^{\boldsymbol{t}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}}}
\end{aligned}
$$

5. b)

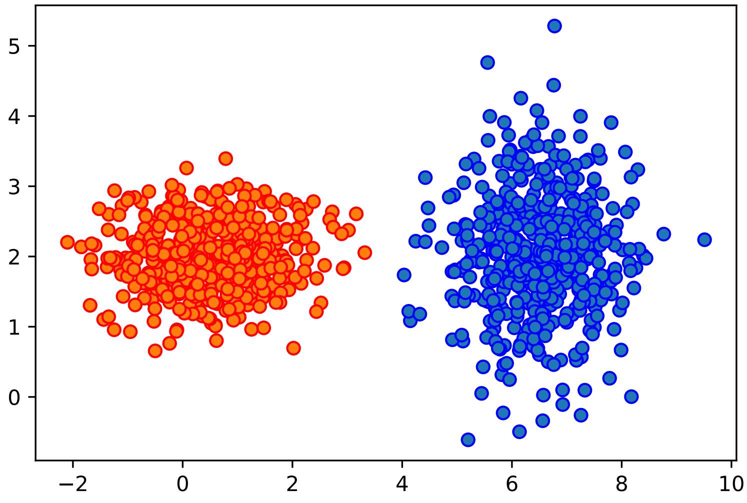

Figure 1: The generated points

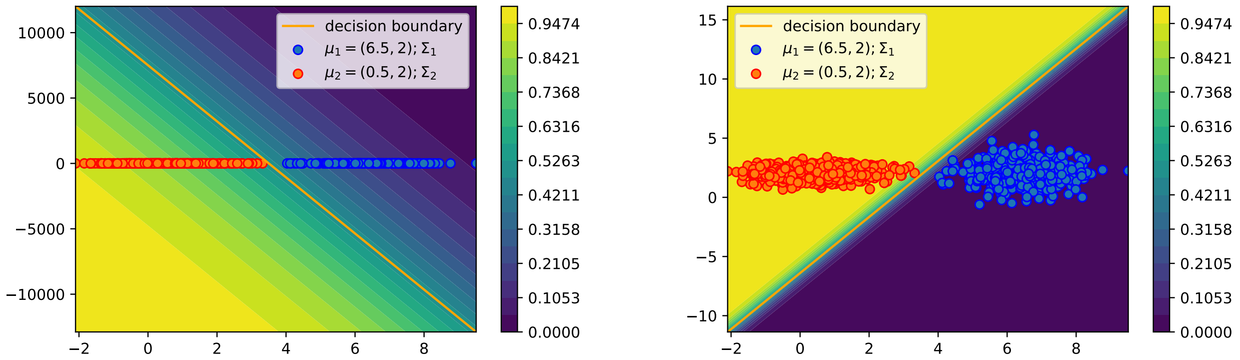

Figure 2: decision boundary and probabilities for $S^2 = 10^{-3}, 10^4$

e,f) The decision boundary is the curve

$$p(t = 1|\boldsymbol{x}) = 0.5 \Leftrightarrow \frac{1}{1 + e^{\tilde{\boldsymbol{w}}^t \tilde{\boldsymbol{x}}}} = 0.5 \Leftrightarrow 1 + e^{\tilde{\boldsymbol{w}}^t \tilde{\boldsymbol{x}}} = 2 \Leftrightarrow \tilde{\boldsymbol{w}}^t \tilde{\boldsymbol{x}} = \log(1) = 0,$$

so it is the line

$$y = -\frac{w_1}{w_2} x - \frac{b}{w_2}.$$

In this case the computed decision boundary really seperates the red and blue training points (for both values of S^2), although in some cases the points may not be linearly separable in two dimensions, which also happened in some tests. The probability around the decision boundary is about 0.5, which is clear, because the boundary was chosen for this reason. If a high variance for $\boldsymbol{w}$ is assumed, the probability decreases rather slowly when moving away from the decision boundary. For higher S^2 the probability decreases much faster and the predictions can be made with more certainty.

6. The accuracy for variing S^2 is

S^2	training	validation
10^{-4}	0.6088	0.6029
10^{-3}	0.8435	0.8377
10^{-2}	0.9003	0.8826
10^{-1}	0.9193	0.9123
1	0.9165	0.9152
10^1	0.9169	0.9109
10^2	0.9165	0.9145
10^3	0.9143	0.9101
10^4	0.9165	0.9138

It fits the previous observation that for higher S^2 the predictions can be made with more certainty, but this effect seems to be limited by $S^2 \approx 0.1$, after that no real improvement is achieved.