Machine Learning
SS 2020
Exercise sheet 2

Solution by
Lorenzo Minecci, Daniel Strenger
11939539, 01531211

May 11, 2020

Exercise 1. (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.

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

5. b)

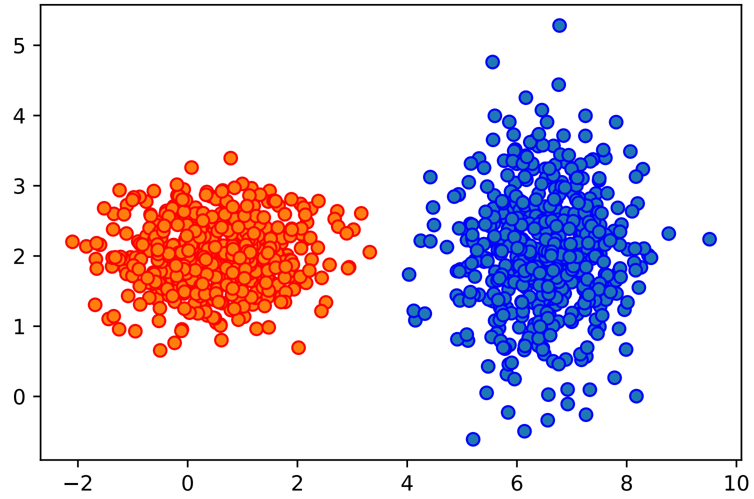

Figure 1: The generated points

e,f) The decision boundary is the curve

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

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

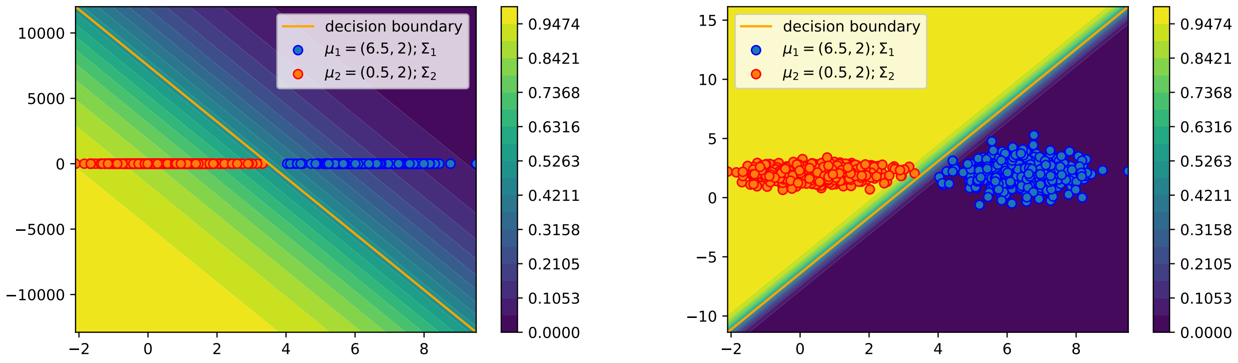

Figure 2: decision boundary and probabilities for $S^2 = 10^{-3}, 10^4$

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.