CS-E3210- Machine Learning Basic Principles
Home Assignment 3 - "Classification"

Your solutions to the following problems should be submitted as one single pdf which does not contain any personal information (student ID or name). The only rule for the layout of your submission is that for each problem there has to be exactly one separate page containing the answer to the problem. You are welcome to use the LaTeX-file underlying this pdf, available under `https://version.aalto.fi/gitlab/junga1/MLBP2017Public`, and fill in your solutions there.

Problem 1: Logistic Regression - I

Answer.

$$
\begin{aligned}
&P(y=1|\mathbf{x};\mathbf{w}) = h^{(\mathbf{w})}(\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} \\
&P(y=-1|\mathbf{x};\mathbf{w}) = 1 - P(y=1|\mathbf{x};\mathbf{w}) = \frac{1}{1+\exp(\mathbf{w}^T\mathbf{x})} \\
&P(y \mid \mathbf{x}) = \frac{1}{1+\exp(-y\mathbf{w}^T\mathbf{x})}
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
&\max_{\mathbf{w}\in\mathbb{R}^2} \ln\big(P(y{=}1|\mathbf{x}^{(1)};\mathbf{w})P(y{=}-1|\mathbf{x}^{(2)};\mathbf{w})\big) \\
=&\max_{\mathbf{w}\in\mathbb{R}^2} \sum_{i=1}^{N=2} \ln P(y^{(i)}|\mathbf{x}^{(i)};\mathbf{w}) \\
=&\max_{\mathbf{w}\in\mathbb{R}^2} \sum_{i=1}^{N=2} \ln\big(\frac{1}{1+\exp(-y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)})}\big) \\
=&\max_{\mathbf{w}\in\mathbb{R}^2} \sum_{i=1}^{N=2} -\ln\big(1+\exp(-y^{(i)}\mathbf{w}^T\mathbf{x}^{(i)})\big) \\
=&\min_{\mathbf{w}\in\mathbb{R}^2} \sum_{i=1}^{N=2} \ln\big(1+\exp\big(-y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)})\big)\big)
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
&\min_{\mathbf{w}\in\mathbb{R}^2} (1/N) \sum_{i=1}^{N=2} L((\mathbf{x}^{(i)},y^{(i)});\mathbf{w}) \\
=&(1/N) \min_{\mathbf{w}\in\mathbb{R}^2} \sum_{i=1}^{N=2} \ln\big(1+\exp\big(-y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)})\big)\big)
\end{aligned} \tag{3}
$$

To get minimal empirical risk, we need to get the minimal $\sum_{i=1}^{N=2} \ln\big(1+\exp\big(-y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)})\big)\big)$. As (2) showed, $\mathbf{w}_{ML}$ satisfies this requirement. So $\mathbf{w}_{ML}$ is a solution to the empirical risk minimization problem.

Problem 2: Logistic Regression - II

Answer.

$$
\begin{aligned}
&\nabla_{\mathbf{w}^{(k)}}(1/N)\sum_{i=1}^{N} L((\mathbf{x}^{(i)}, y^{(i)}); \mathbf{w}^{(k)}) \\
=&(1/N)\nabla_{\mathbf{w}^{(k)}}\sum_{i=1}^{N} \ln\big(1+\exp\big(-y^{(i)}((\mathbf{w}^{(k)})^T\mathbf{x}^{(i)})\big)\big) \\
=&(1/N)\sum_{i=1}^{N} \frac{-y^{(i)}\mathbf{x}^{(i)}\exp\big(-y^{(i)}((\mathbf{w}^{(k)})^T\mathbf{x}^{(i)})\big)}{1+\exp\big(-y^{(i)}((\mathbf{w}^{(k)})^T\mathbf{x}^{(i)})\big)} \\
=&(1/N)\sum_{i=1}^{N} \frac{-y^{(i)}\mathbf{x}^{(i)}}{1+\exp\big(y^{(i)}((\mathbf{w}^{(k)})^T\mathbf{x}^{(i)})\big)}
\end{aligned}
\tag{4}
$$

Problem 3: Bayes' Classifier - I

Answer.

$$p(y \neq h(\mathbf{x})) = 1 - p(y = h(\mathbf{x})) \tag{5}$$

To minimize the error probability, maximizing $p(y = h(\mathbf{x}))$.
Based on MAP,

$$\begin{aligned} h(\mathbf{x}) &= \underset{y \in \{-1,1\}}{\operatorname{argmax}}\, p(y \mid \mathbf{x}) \\ &= \underset{y \in \{-1,1\}}{\operatorname{argmax}}\, p(y) p(\mathbf{x} \mid y) \end{aligned} \tag{6}$$

Expressed in log-space:

$$\begin{aligned} h(\mathbf{x}) &= \underset{y \in \{-1,1\}}{\operatorname{argmax}} \log p(y \mid \mathbf{x}) \\ &= \underset{y \in \{-1,1\}}{\operatorname{argmax}} \big(\log p(y) + \log p(\mathbf{x} \mid y) \big) \end{aligned} \tag{7}$$

Problem 4: Bayes' Classifier - II

Answer.

$$P(y=-1) = 1 - P_1$$

From the Problem 3,

$$\begin{aligned} h(\mathbf{x}) &= \underset{y\in\{-1,1\}}{\operatorname{argmax}} \log p(y \mid \mathbf{x}) \\ &= \underset{y\in\{-1,1\}}{\operatorname{argmax}} \big(\log p(y) + \log p(\mathbf{x} \mid y) \big) \end{aligned} \tag{8}$$

We have had knowledge of the $P(y)$.

$$\underset{y\in\{-1,1\}}{\operatorname{argmax}} \log p(\mathbf{x} \mid y) \tag{9}$$

Based on Maximum Likelihood Method,
Similar to the calculation process in the Homework 1, Problem 3 (The partial derivatives of $\log p(\mathbf{x} \mid y) = 0$), we get

$$\mathbf{m}_s = \tfrac{1}{NP_1} \textstyle\sum_{i=1}^{N} \mathcal{I}(y^{(i)} = 1)\mathbf{x}^{(i)}$$

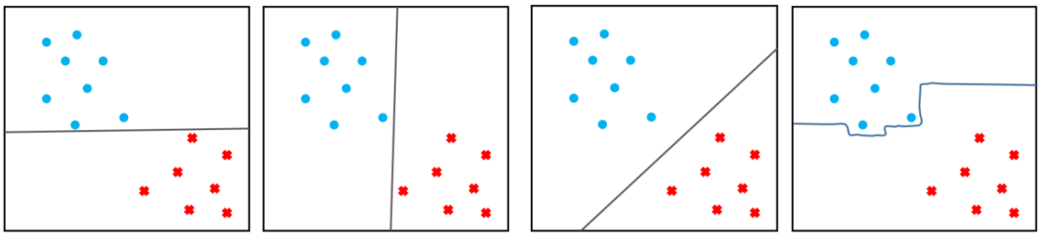
$$\mathbf{m}_w = \tfrac{1}{N(1-P_1)} \textstyle\sum_{i=1}^{N} \mathcal{I}(y^{(i)} = -1)\mathbf{x}^{(i)}$$

$$\mathbf{C}_s = \tfrac{1}{NP_1} \textstyle\sum_{i=1}^{N} \mathcal{I}(y^{(i)} = 1)(\mathbf{x}^{(i)} - \mathbf{m}_s)(\mathbf{x}^{(i)} - \mathbf{m}_s)^T$$

$$\mathbf{C}_w = \tfrac{1}{N(1-P_1)} \textstyle\sum_{i=1}^{N} \mathcal{I}(y^{(i)} = -1)(\mathbf{x}^{(i)} - \mathbf{m}_w)(\mathbf{x}^{(i)} - \mathbf{m}_w)^T$$

Problem 5: Support Vector Classifier

Consider data points with features $\mathbf{x}^{(i)} \in \mathbb{R}^2$ and labels $y^{(i)} \in \{-1, 1\}$. In the figures below, the data points with $y^{(i)} = 1$ are depicted as red crosses and the data points with $y^{(i)} = -1$ are depicted as blue filled circles. Which of the four figures depicts a decision boundary which could have been generated by a SVC. Justify your selection.



(a) (b) (c) (d)

Answer.

Figure (c) depicts a decision boundary which could have been generated by a SVC.

SVC is a linear classifier using hypothesis $h^w(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. It is based on geometry of $\mathbb{X}$ in feature space.
hinge loss:

$$\begin{aligned} L\big((\mathbf{x}, y), h^{\mathbf{w},b}\big) &= \max\{0, 1 - y(\mathbf{w}^T\mathbf{x})\} \\ &= \min \xi \quad s.t. \quad \xi \geq 1 - y(\mathbf{w}^T\mathbf{x}) \end{aligned} \tag{10}$$

Minimizing hinge loss equivalent to maximizing margin $y(\mathbf{w}^T\mathbf{x})$.
There should be some distance between the hyperplane and the support vectors. The sum of the distances of two classes' support vectors from hyperplane is $\gamma = \frac{2}{\|\mathbf{w}\|}$.