# 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

Consider a binary classification problem where the goal is classify or label a webcam snapshot into "winter" ($y = -1$) or "summer" ($y = 1$) based on the feature vector $\mathbf{x} = (x_{\rm g}, 1)^T \in \mathbb{R}^2$ with the image greenness $x_{\rm g}$. A particular classification method is logistic regression, where we classify a datapoint as $\hat{y} = 1$ if $h^{(\mathbf{w})}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) > 1/2$ and $\hat{y} = -1$ otherwise. Here, we used the sigmoid function $\sigma(z) = 1/(1 + \exp(-z))$.

The predictor value $h^{(\mathbf{w})}(\mathbf{x})$ is interpreted as the probability of $y = 1$ given the knowledge of the feature vector $\mathbf{x}$, i.e., $P(y = 1|\mathbf{x}; \mathbf{w}) = h^{(\mathbf{w})}(\mathbf{x})$. Note that the conditional probability $P(y = 1|\mathbf{x}; \mathbf{w})$ is parametrized by the weight vector $\mathbf{w}$. We have only $N = 2$ labeled data points with features $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}$ and labels $y^{(1)} = 1, y^{(2)} = -1$ at our disposal in order to find a good choice for $\mathbf{w}$. Let $\mathbf{w}_{\rm ML}$ be a vector which satisfies

$$P(y{=}1|\mathbf{x}^{(1)}; \mathbf{w}_{\rm ML})P(y{=}{-}1|\mathbf{x}^{(2)}; \mathbf{w}_{\rm ML}) = \max_{\mathbf{w}\in\mathbb{R}^2} P(y{=}1|\mathbf{x}^{(1)}; \mathbf{w})P(y{=}{-}1|\mathbf{x}^{(2)}; \mathbf{w}).$$

Show that the vector $\mathbf{w}_{\rm ML}$ solves the empirical risk minimization problem using logistic loss $L((\mathbf{x}, y); \mathbf{w}) = \ln\big(1 + \exp\big(-y(\mathbf{w}^T\mathbf{x})\big)\big)$, i.e., $\mathbf{w}_{\rm ML}$ is a solution to

$$\min_{\mathbf{w}\in\mathbb{R}^2} (1/N) \sum_{i=1}^{N} L((\mathbf{x}^{(i)}, y^{(i)}); \mathbf{w}).$$

**Answer.**
Because $\min_z \ln(1 + \exp(-z)) = \max_z \ln((1 + \exp(-z))^{-1})$ the logistic loss can be converted to

$$\begin{aligned}
\min_{\mathbf{w}\in\mathbb{R}^2} (1/N) \sum_{i=1}^{N} L((\mathbf{x}^{(1)}, y^{i}); \mathbf{w}) &= \max_{\mathbf{w}\in\mathbb{R}^2} \sum_{i=1}^{N} \ln\Big(\frac{1}{1 + \exp(-y^i(\mathbf{w}\mathbf{x}))}\Big) \\
&= \max_{\mathbf{w}\in\mathbb{R}^2} \left( \sum_{y^i=1} \ln\Big(\frac{1}{1 + \exp(-\mathbf{w}\mathbf{x}^{(y^i)})}\Big) + \sum_{y^i=-1} \ln\Big(\frac{1}{1 + \exp(\mathbf{w}\mathbf{x}^{(y^i)})}\Big) \right) \\
&= \max_{\mathbf{w}\in\mathbb{R}^2} \left( \sum_{y^i=1} \ln\Big(\frac{1}{1 + \exp(-\mathbf{w}\mathbf{x}^{(y^i)})}\Big) + \sum_{y^i=-1} \ln\Big(1 - \frac{1}{\exp(-\mathbf{w}\mathbf{x}^{(y^i)})}\Big) \right) \\
&= \max_{\mathbf{w}\in\mathbb{R}^2} \left( \ln\Big(\frac{1}{1 + \exp(-\mathbf{w}\mathbf{x}^{(1)})}\Big) + \ln\Big(1 - \frac{1}{1 + \exp(-\mathbf{w}\mathbf{x}^{(2)})}\Big) \right) \\
&= \max_{\mathbf{w}\in\mathbb{R}^2} \Big( \ln(P(y = 1|\mathbf{x}^{(1)}; \mathbf{w}) + \ln(P(y = -1|\mathbf{x}^{(2)}; \mathbf{w}) \Big) \\
&= \max_{\mathbf{w}\in\mathbb{R}^2} \ln(P(y = 1|\mathbf{x}^{(1)}; \mathbf{w})P(y = -1|\mathbf{x}^{(2)}; \mathbf{w})) \\
&= \max_{\mathbf{w}\in\mathbb{R}^2} P(y = 1|\mathbf{x}^{(1)}; \mathbf{w})P(y = -1|\mathbf{x}^{(2)}; \mathbf{w}),
\end{aligned}$$

and thus the vector $\mathbf{w}_{\rm ML}$ is the solution to logistic loss. In above the fact $P(y = -1|\mathbf{x}^{(2)}; \mathbf{w}) = 1 - P(y = 1|\mathbf{x}^{(2)}; \mathbf{w})$ was used.

## Problem 2: Logistic Regression - II

Consider a binary classification problem where the goal is classify or label a webcam snapshot into "winter" ($y = -1$) or "summer" ($y = 1$) based on the feature vector $\mathbf{x} = (x_{\mathrm{g}}, 1)^T \in \mathbb{R}^2$ with the image greenness $x_{\mathrm{g}}$. A particular classification method is logistic regression, where we classify a datapoint as $\hat{y} = 1$ if $h^{(\mathbf{w})}(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) > 1/2$ and $\hat{y} = -1$ otherwise. Here, we used the sigmoid function $\sigma(z) = 1/(1 + \exp(-z))$.

Given some labeled snapshots $\mathbb{X} = \left\{(x^{i)}, y^{(i)})\right\}_{i=1}^{N}$, we choose the weight vector $\mathbf{w}$ by empirical risk minimization using logistic loss $L((\mathbf{x}, y); \mathbf{w}) = \ln\left(1+\exp\left(-y(\mathbf{w}^T\mathbf{x})\right)\right)$, i.e.,

$$\mathbf{w}_{\mathrm{opt}} = \arg\min_{\mathbf{w}\in\mathbb{R}^2} \underbrace{(1/N)\sum_{i=1}^{N} L((\mathbf{x}^{(i)}, y^{(i)}); \mathbf{w})}_{=f(\mathbf{w})}. \tag{1}$$

Since there is no simple closed-form expression for $\mathbf{w}_{\mathrm{opt}}$, we have to use some optimization method for (approximately) finding $\mathbf{w}_{\mathrm{opt}}$. One extremely useful such method is gradient descent which starts with some initial guess $\mathbf{w}^{(0)}$ and iterates

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha\nabla f(\mathbf{w}^{(k)}), \tag{2}$$

for $k = 0, 1, \ldots$. For a suitably chosen step-size $\alpha > 0$ one can show that $\lim_{k\to\infty} \mathbf{w}^{(k)} = \mathbf{w}_{\mathrm{opt}}$. Can you find a simple closed-form expression for the gradient $\nabla f(\mathbf{w}^{(k)})$ in terms of the current iterate $\mathbf{w}^{(k)}$ and the data points $\mathbb{X} = \left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^{N}$.

**Answer.**

By marking $t^i = \max(0, y^i)$ we can write the equation (1) as

$$\max_{\mathbf{w}\in\mathbb{R}^2}\left(\sum_{y^i=1}\ln(p_{y^i}) + \sum_{y^i=-1}\ln(1-p_{y^i})\right) \;=\; \min_{\mathbf{w}\in\mathbb{R}^2} -\sum_{i}^{N}\left(t^i\ln(p_{y^i}) + (1-t^i)\ln(1-p_{y^i})\right), \tag{3}$$

where $p_{y^i}$ was defined as $p_{y^i} = \frac{1}{1+\exp(-\mathbf{w}\mathbf{x}^i)}$. The derivative of $p_{y^i}$ with respect to $w_j$ can be written as $\frac{\partial p_{y^i}}{\partial w_j} = \frac{\partial p_{y^i}}{\partial(\mathbf{w}\mathbf{x})}\frac{\partial(\mathbf{w}\mathbf{x})}{\partial w_j} = p_{y^i}(1-p_{y^i})x_j^i$. Thus taking derivative of equation (3) with respect to $w_j$ we get for $\nabla f(w_j^k)$

$$\begin{aligned}
\nabla f(w_j^k) &= -\sum_i\left(t^i\frac{1}{p_{y^i}}(1-p_{y^i})p_{y^i}x_j + (1-t^i)\frac{1}{1-p_{y^i}}(-1)p_{y^i}(1-p_{y^i})x_j\right)\\
&= -\sum_i\left(t^i - t^ip_{y^i} - p_{y^i} + t^ip_{y^i}\right)x_j\\
&= -\sum_i(t^i - p_{y^i})x_j. \qquad (4)
\end{aligned}$$

Now the $\nabla f(w_j^k)$'s can be written in vector form as

$$\nabla f(\mathbf{w}^k) = -\left(\max(\mathbf{0}, \mathbf{y}) - \frac{1}{1+\exp(-\mathbf{w}^{k-1}X^T)}\right)^T X, \tag{5}$$

where $\max(\mathbf{0}, \mathbf{y})$ is a vector of length $N$ for which the $i$-th element is zero if $y_i = -1$ and one if $y_i = 1$. The matrix $X$ is a matrix which rows are the feature vectors $\mathbf{x}^i$.

## Problem 3: Bayes' Classifier - I

Consider a binary classification problem where the goal is classify or label a webcam snapshot into "winter" ($y = -1$) or "summer" ($y = 1$) based on the feature vector $\mathbf{x} = (x_{\rm g}, 1)^T \in \mathbb{R}^2$ with the image greenness $x_{\rm g}$. We might interpret the feature vector and label as (realizations) of random variables, whose statistics is specified by a joint distribution $p(\mathbf{x}, y)$. This joint distribution factors as $p(\mathbf{x}, y) = p(\mathbf{x}|y)p(y)$ with the conditional distribution $p(\mathbf{x}|y)$ of the feature vector given the true label $y$ and the prior distribution $p(y)$ of the label values. The prior probability $p(y = 1)$ is the fraction of overall summer snapshots. Assume that we know the distributions $p(\mathbf{x}|y)$ and $p(y)$ and we want to construct a classifier $h(\mathbf{x})$, which classifies a snapshot with feature vector $\mathbf{x}$ as $\hat{y} = h(\mathbf{x}) \in \{-1, 1\}$. Which classifier map $h(\cdot) : \mathbf{x} \mapsto \hat{y} = h(\mathbf{x})$, mapping the feature vector $\mathbf{x}$ to a predicted label $\hat{y}$, yields the smallest error probability (which is $p(y \neq h(\mathbf{x}))$) ?

**Answer.**

The $p(\mathbf{x}, y)$ can be also written as $p(\mathbf{x}, y) = p(y|\mathbf{x})p(\mathbf{x})$ and thus we get for the posterior $p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}$. Now, because the distribution $p(\mathbf{x}|y)$ was known exactly, the classifier $\hat{y} = h(\mathbf{x})$ that yields the smallest error probability is

$$\hat{y} = \underset{y \in \{-1,1\}}{\operatorname{argmax}} \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}. \tag{6}$$

## Problem 4: Bayes' Classifier - II

Reconsider the binary classification problem of Problem 3, where the goal is classify or label a webcam snapshot into "winter" ($y = -1$) or "summer" ($y = 1$) based on the feature vector $\mathbf{x} = (x_{\rm g}, 1)^T \in \mathbb{R}^2$ with the image greenness $x_{\rm g}$. While in Problem 3 we assumed perfect knowledge of the joint distribution $p(\mathbf{x}, y)$ of features $\mathbf{x}$ and label $y$ (which are modelled as random variables), now we consider only knowledge of the prior probability $P(y = 1)$, which we denote $P_1$. A useful "guess" for the distribution of the features $\mathbf{x}$, given the label $y$, is via a Gaussian distribution. Thus, we assume

$$p(\mathbf{x}|y = 1; \mathbf{m}_s, \mathbf{C}_s) = \frac{1}{\sqrt{\det\{2\pi\mathbf{C}_s\}}} \exp(-(1/2)(\mathbf{x}-\mathbf{m}_s)^T\mathbf{C}_s^{-1}(\mathbf{x}-\mathbf{m}_s))$$

and, similarly,

$$p(\mathbf{x}|y = -1; \mathbf{m}_w, \mathbf{C}_w) = \frac{1}{\sqrt{\det\{2\pi\mathbf{C}_w\}}} \exp(-(1/2)(\mathbf{x}-\mathbf{m}_w)^T\mathbf{C}_w^{-1}(\mathbf{x}-\mathbf{m}_w)).$$

How would you choose (fit) the parameters $\mathbf{m}_s, \mathbf{m}_w \in \mathbb{R}^2$ and $\mathbf{C}_s, \mathbf{C}_w \in \mathbb{R}^{2\times 2}$ for (to) a given labeled dataset $\mathbb{X} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$.

**Answer.**

Because the samples $x^{(i)}, y^{(i)}$ are independent the parameters $\mathbf{m}_s, \mathbf{m}_w, \mathbf{C}_s$ and $\mathbf{C}_w$ can be obtained by maximizing the respective log-likelihood function with respect to the parameter. For parameters $\mathbf{m}_s, \mathbf{C}_s$ the log-likelihood can be written as

$$\begin{aligned}\mathcal{L}_{y=1} &= \sum_{y^i=1} \log(P(\mathbf{x}^i|y = 1; \mathbf{m}_s, \mathbf{C}_s)) \\ &= \sum_{y^i=1} \left(-\frac{1}{2}\log(2\pi^n) - \frac{1}{2}\log(\det(\mathbf{C}_s)) - \frac{1}{2}(\mathbf{x} - \mathbf{m}_s)^T\mathbf{C}_s^{-1}(\mathbf{x} - \mathbf{m}_s)\right). \qquad (7)\end{aligned}$$

Now the optimal $\mathbf{m}_s$ is obtained by setting the derivative of $\mathcal{L}_{y=1}$ with respect to $\mathbf{m}_s$ to zero, *i.e.*,

$$\frac{\partial\mathcal{L}_{y=1}}{\partial\mathbf{m}_s} = \mathbf{C}_s^{-1}\sum_{y^i=1}(\mathbf{x}^i - \mathbf{m}_s). \qquad (8)$$

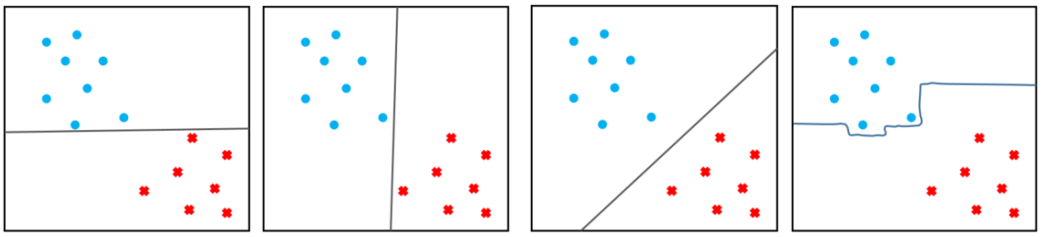
Setting eq. (8) to zero and solving $\mathbf{m}_s$ gives $\mathbf{m}_s = \frac{\sum_{y=1}\mathbf{x}^i}{N_{y=1}}$. Similarly the $\mathbf{C}_s$ is obtained by setting the derivative of $\mathcal{L}_{y=1}$ with respect to $\mathbf{C}_s$ to zero giving

$$\frac{\partial\mathcal{L}_{y=1}}{\partial\mathbf{C}_s} = \sum_{y^i=1}\left(-\frac{1}{2}\mathbf{C}_s^{(-1)} + \frac{1}{2}\mathbf{C}_s^{(-1)}(\mathbf{x} - \mathbf{m}_s)(\mathbf{x} - \mathbf{m}_s)^T\mathbf{C}_s^{-1}\right). \qquad (9)$$

Setting eq. (9) to zero and solving $\mathbf{C}_s$ gives $\mathbf{C}_s = \frac{1}{N}\sum_{y^i=1}(\mathbf{x} - \mathbf{m}_s)(\mathbf{x} - \mathbf{m}_s)^T$. In equations (8) and (9) the identieties $\frac{\partial(\mathbf{x}-\mathbf{s})^T\mathbf{W}(\mathbf{x}-\mathbf{s})}{\partial\mathbf{s}} = 2\mathbf{W}(\mathbf{x} - \mathbf{s})$, $\frac{\partial\ln|\det(\mathbf{X})|}{\partial\mathbf{X}} = (\mathbf{X}^T)^{-1}$ and $\frac{\partial\mathbf{a}^T\mathbf{X}^{-1}\mathbf{b}}{\partial\mathbf{X}} = -\mathbf{X}^{-T}\mathbf{a}\mathbf{b}^T\mathbf{X}^{-T}$ from matrix cook-book were used. For $\mathbf{m}_w$ and $\mathbf{C}_w$ are obtained similarly by replacing the summation over $y^i = 1$ to $y^i = -1$.

## 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.**
The figures (a) and (b) are clearly not produced by SVC. That is because a linear classifier is clearly used and the linear SVC maximizes the marginal, *i.e.*, linear SVC produces a plane that is at maximum distance from both label groups, which cleary is not the case in (a) and (b). However in plot (c) the separating plane is clearly at maximum distance from both label sets and is thus produced by SVC algorithm. The plot (c) is not generated by linear classifier. That is why it might possible that it is generated by non-linear SVD??