Machine Learning 1 — Mock Exam WS 2017 / 2018

1 Probability Theory

Problem 1 [0 point] You have two coins, C_1 and C_2. Let the outcome of a coin toss be either *heads* ($C_i = 1$) or *tails* ($C_i = 0$) for $i = 1, 2$. C_1 is a fair coin. However, C_2 depends on C_1: If C_1 shows *heads* ($C_1 = 1$), C_2 will show *heads* with probability 0.7. If C_1 shows *tails* ($C_1 = 0$), C_2 will show *heads* with probability 0.5. Now you toss C_1 and C_2 in sequence once. You observe the sum of the two coins $S = C_1 + C_2 = 1$. What is the probability that C_1 shows *tails* and C_2 shows *heads*?

2 Parameter Inference / Full Bayesian Approach

For a Naive Bayes classifier we assume the following model:

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}|\Theta) &= p(\mathbf{x}|\mathbf{y}, \Theta)p(\mathbf{y}|\Theta) \\ &= p(\mathbf{x}|\mathbf{y}, \theta, \pi)p(\mathbf{y}|\theta, \pi) \\ &= p(\mathbf{x}|\mathbf{y}, \theta)p(\mathbf{y}|\pi) \\ &= \prod_{v=1}^{V} p(x_v|\mathbf{y}, \theta)p(\mathbf{y}|\pi) \\ &= \prod_{c=1}^{C}\prod_{v=1}^{V} p(x_v|\theta_{vc})^{y_c} \prod_{c'=1}^{C} \pi_{c'}^{y_{c'}} \end{aligned}$$

where we leave open, which model for the class-conditional densities $p(x_v|\theta_{vc})$ we are using.

Problem 2 [0 point] For this model, write down the posterior distribution for the parameters $p(\Theta|\mathcal{D})$, where $\mathcal{D} = \{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}\}_{n=1}^{N}$! It suffices to specify $p(\Theta|\mathcal{D})$ on the $\propto$ level (that is up to constants in Θ) and name the distributions you are introducing as far as the model specification goes.

Problem 3 [0 point] Show that for the full Bayesian estimation of the class $\mathbf{y}$ for a new data point $\mathbf{x}$ we have

$$p(y_c = 1|\mathbf{x}, \mathcal{D}) \propto \int \prod_{v=1}^{V} p(x_v|\theta_{vc})p(\theta|\mathcal{D})d\theta \int \pi_c p(\pi|\mathcal{D})d\pi$$

3 Regularized Logistic Regression

We employ a logistic regression model to classify the data which are plotted in the below figure,

$$\mathbf{p}(y = 1 \mid \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1 + \exp(\mathbf{w}^T\mathbf{x})}.$$

We fit the data by the maximum likelihood approach, and minimise the negative log-likelihood $-l(\mathbf{w})$, thus the objective function is

$$J(\mathbf{w}) = -l(\mathbf{w}).$$

Problem 1 [0 point] You have two coins, C_1 and C_2. Let the outcome of a coin toss be either *heads* ($C_i = 1$) or *tails* ($C_i = 0$) for $i = 1, 2$. C_1 is a fair coin. However, C_2 depends on C_1. If C_1 shows *heads* ($C_1 = 1$), C_2 will show *heads* with probability 0.7. If C_1 shows *tails* ($C_1 = 0$), C_2 will show *heads* with probability 0.5. Now you toss C_1 and C_2 in sequence once. You observe the sum of the two coins $S = C_1 + C_2 = 1$. What is the probability that C_1 shows *tails* and C_2 shows *heads*?

$p(C_2 = 1 \mid C_1 = 1) = 0.7$

$p(C_2 = 1 \mid C_1 = 0) = 0.5$

$p(C_1 = 0, C_2 = 1 \mid S = 1)$

$$= \frac{p(S=1 \mid C_1=0, C_2=1)\; p(C_1=0, C_2=1)}{p(S=1)}$$

$$= \frac{p(S=1 \mid C_1=0, C_2=1)\; p(C_2=1 \mid C_1=0)\; p(C_1=0)}{p(S=1)}$$

$$= p(S=1 \mid C_1=0, C_2=0)\, p(C_2=0 \mid C_1=0)\, p(C_1=0)$$
$$p(S=1 \mid C_1=0, C_2=1)\, p(C_2=1 \mid C_1=0)\, p(C_1=0)$$
$$p(S=1 \mid C_1=1, C_2=0)\, p(C_2=0 \mid C_1=1)\, p(C_1=1)$$
$$p(S=1 \mid C_1=1, C_2=1)\, p(C_2=1 \mid C_1=1)\, p(C_1=1)$$

$$= 1 \cdot 0.5 \cdot \frac{1}{2} + 1 \cdot 0.3 \cdot \frac{1}{2} = \frac{1}{4} + \frac{3}{20} = \frac{8}{20} = \frac{2}{5}$$

$$= \frac{1 \cdot \frac{1}{2} \cdot \frac{1}{2}}{\frac{2}{5}} = \frac{5}{8}$$

For a Naive Bayes classifier we assume the following model:

$$\begin{aligned} p(\mathbf{x}, \mathbf{y} | \Theta) &= p(\mathbf{x} | \mathbf{y}, \Theta) p(\mathbf{y} | \Theta) \\ &= p(\mathbf{x} | \mathbf{y}, \theta, \pi) p(\mathbf{y} | \theta, \pi) \\ &= p(\mathbf{x} | \mathbf{y}, \theta) p(\mathbf{y} | \pi) \\ &= \prod_{v=1}^{V} p(x_v | \mathbf{y}, \theta) p(\mathbf{y} | \pi) \\ &= \prod_{c=1}^{C} \prod_{v=1}^{V} p(x_v | \theta_{vc})^{y_c} \prod_{c'=1}^{C} \pi_{c'}^{y_{c'}} \end{aligned}$$

where we leave open, which model for the class-conditional densities $p(x_v|\theta_{vc})$ we are using.

Problem 2 [0 point] For this model, write down the posterior distribution for the parameters $p(\Theta|\mathcal{D})$, where $\mathcal{D} = \{\mathbf{x}^{(n)}, \mathbf{y}^{(n)}\}_{n=1}^{N}$! It suffices to specify $p(\Theta|\mathcal{D})$ on the $\propto$ level (that is up to constants in Θ) and name the distributions you are introducing as far as the model specification goes.

Problem 3 [0 point] Show that for the full Bayesian estimation of the class $\mathbf{y}$ for a new data point $\mathbf{x}$ we have

$$p(y_c = 1 | \mathbf{x}, \mathcal{D}) \propto \int \prod_{v=1}^{V} p(x_v | \theta_{vc}) p(\theta | \mathcal{D}) d\theta \int \pi_c \underbrace{p(\pi | \mathcal{D})}_{p(y|\pi)} d\pi$$

$$p(\theta|D) \propto p(D|\theta)\cdot p(\theta)$$

$$\propto \prod_{n=1}^{N} \sum_{c=1}^{C} \prod_{v=1}^{V} p(x_v|\theta_{vc})^{y_c} \prod_{c'=1}^{C'} \pi_c^{y_c} \cdot p(\theta) = p(\theta|\alpha,\beta) \propto p(\theta|\beta)(\pi|\alpha)$$

$$p(y_c|x,D) = \int p(y_c,\theta|x,D)\,d\theta$$

$$= \int p(y_c|\theta,x)\,p(\theta|D)\,d\theta$$

$$\propto \int p(x|y_c,\theta)\cdot p(y_c|\theta)\cdot p(\theta|D)\,d\theta$$

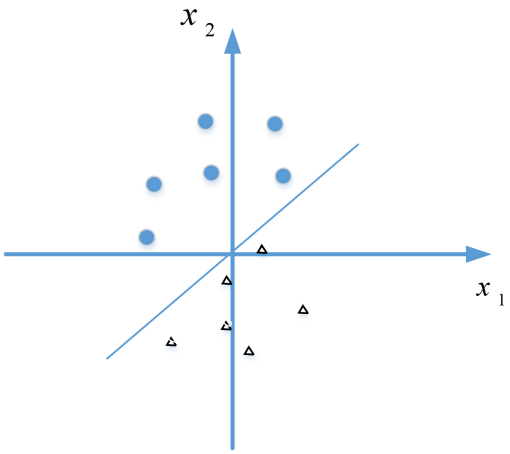
$$= \frac{p(x|y_c,\theta)\cdot p(y_c|\theta)}{p(y_c)}$$

$$\propto \int p(x|y_c,\theta,\pi)\,p(y_c|\theta,\pi)\,p(\theta|D)\,p(\pi|D)\,d\theta\,d\pi$$

$$\propto \int p(x|y_c,\theta)\,p(\theta|D)\,d\theta \int \pi_c\,p(\pi|D)\,d\pi$$

$$\propto \int \prod_{v=1}^{V} p(x_v|\theta_{vc})\,p(\theta|D)\,d\theta \int \pi_c\,p(\pi|D)\,d\pi$$

We get the decision boundary as shown in the figure with zero misclassification error.



Problem 4 [0 point] Now, we regularise w_2 and minimise

$$J_0(\mathbf{w}) = -l(\mathbf{w}) + \lambda w_2^2.$$

Draw the area that the decision boundary can be in and explain your work.

4 Kernels

The following information about kernels *might* be helpful.

Let K_1 and K_2 be kernels on $\mathcal{X} \subseteq \mathbb{R}^n$, then the following functions are kernels:

1. $K(\boldsymbol{x}, \boldsymbol{y}) = K_1(\boldsymbol{x}, \boldsymbol{y}) + K_2(\boldsymbol{x}, \boldsymbol{y})$
2. $K(\boldsymbol{x}, \boldsymbol{y}) = \alpha\, K_1(\boldsymbol{x}, \boldsymbol{y})$ for $\alpha > 0$
3. $K(\boldsymbol{x}, \boldsymbol{y}) = K_1(\boldsymbol{x}, \boldsymbol{y})\, K_2(\boldsymbol{x}, \boldsymbol{y})$
4. $K(\boldsymbol{x}, \boldsymbol{y}) = K_3(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}))$ for K_3 kernel on $\mathbb{R}^m$ and $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^M$
5. $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T B \boldsymbol{y}$ for $B \in \mathbb{R}^{n \times n}$ symmetric and positive semi-definite

Problem 5 [0 point] We have $\mathbf{x} = [x_1\ x_2]^T$. Given the mapping

$$\varphi(x) = [1\ \ x_1^2\ \ \sqrt{2}x_1x_2\ \ x_2^2\ \ \sqrt{2}x_1\ \ \sqrt{2}x_2]^T$$

Determine the kernel $K(\mathbf{x}, \mathbf{y})$. Simplify your answer.

Problem 6 [0 point] Let Z be a set of *finite* size. Show that the function

$$K_0(X, Y) = |X \cap Y|$$

is a valid kernel, provided that $X \subseteq Z$ and $Y \subseteq Z$. Remember that Z is finite, i.e. $Z = \{z_1, z_2, \ldots, z_N\}$.

Problem 7 [0 point] Again, let Z be a set of *finite* size. Show that the function

$$K(X, Y) = 2^{|X \cap Y|}$$

is a valid kernel, provided that $X \subseteq Z$ and $Y \subseteq Z$.

Even if you did not succeed in the previous exercise, you may assume that $K_0(X, Y)$ is a valid kernel.

5 Neural networks

Problem 8 [0 point] Geoffrey has a data set with input $\mathbf{x} \in \mathbb{R}^2$ and output $y \in \mathbb{R}^1$. He tests a neural network A with one hidden layer and 9 neurons in that layer (not counting the bias of that layer as a node). He also tests a neural network B with two hidden layers and three neurons for each of these layers (again not counting the biases as nodes). How many free parameters do the two models have? Show your calculation!.

Problem 9 [0 point] Consider a neural network for regression with one output neuron. For that case, the model would be

$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|y^{NN}(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

where $y^{NN}(\mathbf{x}, \mathbf{w})$ denotes the output of the neural network.

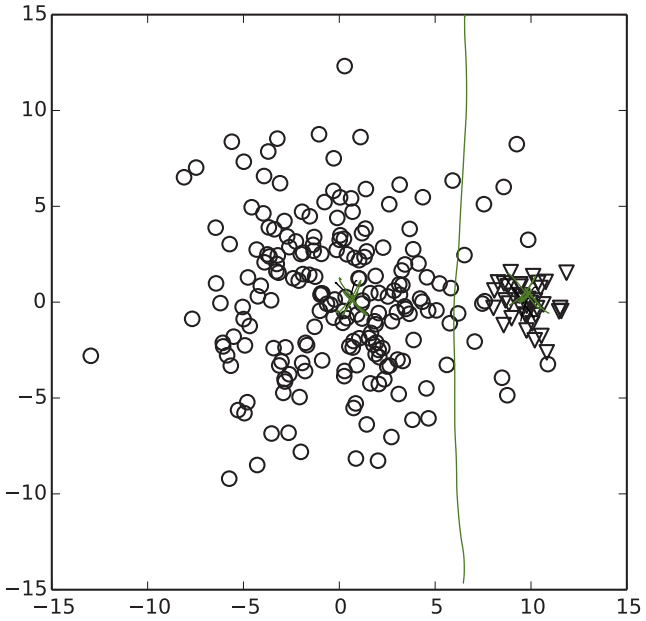
Show that

$$\delta = \frac{\partial E_n}{\partial a} = y^{NN}(\mathbf{x}^{(n)}, \mathbf{w}) - y^{(n)}.$$

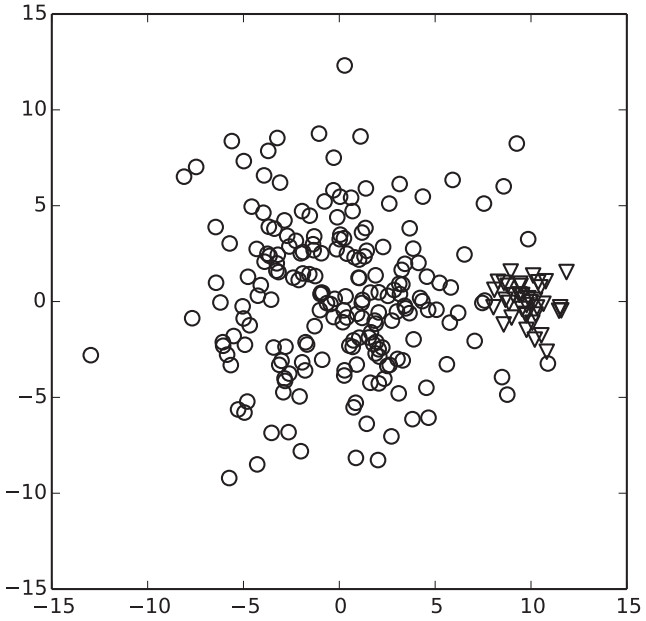
Which activation function will you have to use at the output neuron to arrive at this result?

6 Clustering

Problem 10 [0 point] Consider the plot below. The data is assumed to have been sampled from two different class-conditional densities and the corresponding class labels are indicated with circles (200 data points) and triangles (40 data points). Now assume that you are given the data of the plot without the class labels. In the plot, draw the resulting decision boundary for cluster assignments for a converged run of k-means (Loyd's algorithm) with two centroids.



Problem 11 [0 point] How could we define an analogous hard decision boundary for cluster assignments if instead of k-means (Loyd's algorithm) we would use the EM algorithm with a Gaussian mixture model with two components and individual full covariance matrices as clustering approach? Draw a likely decision boundary qualitatively in the figure!



Problem 12 [0 point] Describe the main steps of the EM algorithm applied to a Gaussian mixture model.

7 Linear Regression

Problem 13 [0 point] For ridge regression, we have the following well known objective function:

$$J(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1})^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}) + \lambda\mathbf{w}^T\mathbf{w}$$

where $\mathbf{1} = (1, 1, \ldots, 1)^T$ and where, in contrast to the lecture slides, we have NOT "absorbed" w_0 into $\mathbf{w}$ by padding each $\mathbf{x}$ with an additional component $= 1$

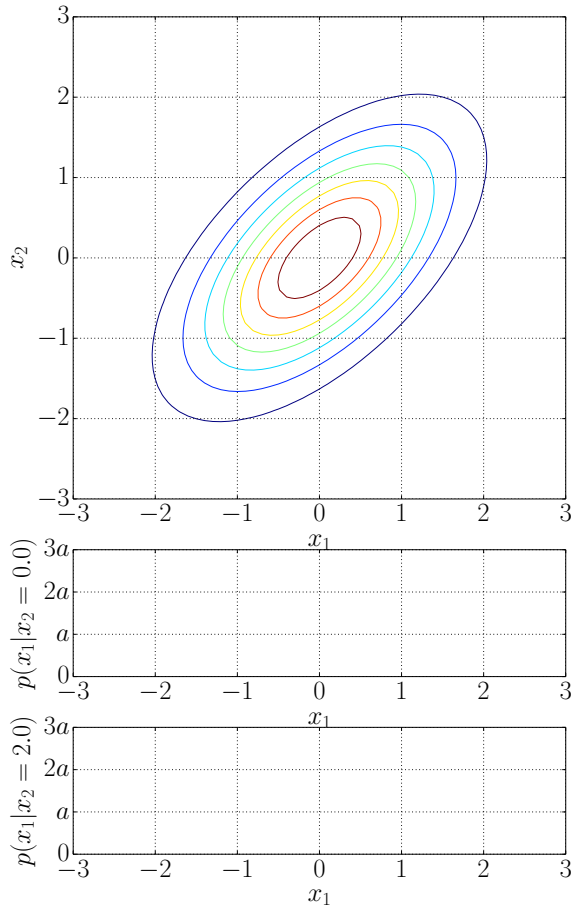
Assuming $\bar{\mathbf{x}} = 0$, derive the expression for the optimizer for $\mathbf{w}$:

$$\hat{\mathbf{w}}_{ridge} = (\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

8 Multivariate Gaussian

Problem 14 [0 point] The plot below shows a joint Gaussian distribution $p(x_1, x_2)$. Qualitatively draw the conditionals $p(x_1|x_2 = 0)$ and $p(x_1|x_2 = 2)$ in the given coordinate systems (In the coordinate systems, the vertical axes have an arbitrary scale factor a to avoid having to deal with exact numbers for the vertical axes' values).

Hint: for a general multivariate Gaussian $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma})$, where $\mathbf{x} \in \mathbb{R}^D$, the conditional $p(\mathbf{x}_1|\mathbf{x}_2)$ (where we split $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)^T$ into $\mathbf{x}_1 \in \mathbb{R}^M$ and $\mathbf{x}_1 \in \mathbb{R}^{D-M}$) is given by $p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\mu_{1|2}, \mathbf{\Sigma}_{1|2})$ with $\mu_{1|2} = \mu_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \mu_2)$ and $\mathbf{\Sigma}_{1|2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$, where $\mu = \begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}$ and $\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$



9 Constrained optimization

Find the box with the maximum volume which has surface area no more than $S \in \mathbb{R}^+$.

Problem 15 [0 point] Derive the Lagrangian of the problem and the corresponding Lagrange dual function. Hint: set the parameters of the length, width and height to be l, w, h respectively.

Problem 16 [0 point] Solve the dual problem and give the solution to the original problem. You may assume without proof that the duality gap is zero.

10 Variational Inference

Problem 17 [0 point] Show that evidence lower bound (ELBO), defined as

$$\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right]$$

is a lower bound to the evidence

$$\log p(\mathbf{x}).$$