Exam (8 April 2021) Attempt review

Machine Learning 1 (Technische Universität Berlin)

Dashboard ▸ My courses ▸ [WiSe 20/21] ML1 ▸ General ▸ Exam (8 April 2021)

Started on	Thursday, 8 April 2021, 8:30 AM
State	Finished
Completed on	Thursday, 8 April 2021, 10:30 AM
Time taken	1 hour 59 mins
Grade	**50.00** out of 100.00

Information

Eigenständigkeitserklärung / Declaration of Independence

By proceeding further, you confirm that you are completing the exam alone, without other resources than those that are authorized. Authorized resources include the ML1 course material, personal notes, online APIs documentation, and calculation/plotting tools.

Question **1**
Incorrect
Mark 0.00 out of 5.00

Which of the following is **True**: A Gaussian Process (GP):

○ a. defines a multivariate Gaussian distribution over output variables, with covariance determined by input similarity.

○ b. defines a multivariate Gaussian distribution over input variables, with covariance determined by output similarity.

● c. defines a multivariate distribution over output variables, with input drawn from a Gaussian distribution. ✗

○ d. defines a multivariate Gaussian distribution over input variables.

Your answer is incorrect.

The correct answer is:
defines a multivariate Gaussian distribution over output variables, with covariance determined by input similarity.

Question **2**
Correct
Mark 5.00 out of 5.00

Which of the following is **True**: In learning theory, the VC (Vapnik-Chervonenkis) bound:

○ a. Is an upper-bound to the generalization error of a trained ML classifier of any complexity.

○ b. Is a lower-bound to the generalization error of a trained ML classifier of any complexity.

● c. Is an upper-bound to the generalization error of a trained ML classifier of limited complexity. ✓

○ d. Is a lower-bound to the generalization error of a trained ML classifier of limited complexity.

Your answer is correct.

The correct answer is:
Is an upper-bound to the generalization error of a trained ML classifier of limited complexity.

Question **3**

Correct

Mark 5.00 out of 5.00

Which of the following is **True**: k-means:

○ a. Is a supervised learning algorithm similar to k-nearest neighbors.

○ b. Has a convex objective and always converges to the global optimum.

● c. Learns a solution that depends on the initialization. ✓

○ d. Is a supervised learning algorithm for representation learning.

Your answer is correct.

The correct answer is:
Learns a solution that depends on the initialization.

Question **4**

Incorrect

Mark 0.00 out of 5.00

Which of the following is **True**: A Product of Experts:

○ a. Is an extension of a mixture model where each mixture element is forced to be Gaussian.

○ b. Is an extension of a mixture model where each mixture element can be Gaussian with non-isotropic covariance.

○ c. Learns less local features than a mixture model.

● d. Is an extension of a mixture model where each mixture element can be non-Gaussian with isotropic covariance. ✗

Your answer is incorrect.

The correct answer is:
Learns less local features than a mixture model.

Information

Assume you would like to build a neural network that implements some function $f : \mathbb{R}_+^d \to \mathbb{R}$ mapping inputs assumed to be positive to a real-valued output. For this, you have at your disposal neurons of the type

$$a_j = \max\left(0, \textstyle\sum_i a_i w_{ij} + b_j\right)$$

where $\sum_i$ sums over the indices of the incoming neurons, and zero otherwise. Denote by a_1 and a_2 the two input neurons (initialized to the value x_1 and x_2 respectively and which are always positive). Denote by a_3, a_4 the hidden neurons, and by a_5 the output neuron.

Question **5**

Complete

Mark 0.00 out of 5.00

Give the weights and biases associated to a neural network with the structure above and that implements the function $f(x) = \max(x_1, x_2)$.

w13 = 0.5, w23 = -0.5, b3 = 1
w14 = -0.5 , w24 = 0.5 , b4 = 1

w35 = 1 , w45 = 1, b5 = (x1+x2)/2

Comment:

Question **6**

Complete

Mark 5.00 out of 5.00

Explain what would be the minimum number of required hidden neurons if not taking two dimensions as input, but d dimensions, and replacing x_1, x_2 by $x_1, x_2, \ldots, x_d$ in the formula above. In this exercise, you can consider architectures of any depth, where neurons can be connected in an arbitrary fashion, but where the nonlinear activation function is of the type mentioned above.

We would need d hidden neurons. Each input dimension requires one hidden neuron and every input variable is connected to each hidden layer. There is no need to build a more complex network with increasing dimensions.

Comment:

Question **7**

Complete

Mark 5.00 out of 5.00

Assume you observe the data point $\boldsymbol{x} = (2, 3)$. Give the value of the partial derivative $\partial a_5 / \partial x_1$ for this data point.

0

Comment:

Information

A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is positive semi-definite (PSD) if for any sequence of N data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \mathbb{R}^d$ and real-valued scalars $c_1, \ldots, c_N$, the following inequality holds:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$$

In the ML1 course, various kernels have been shown to be PSD, for example, the linear kernel is PSD, a sum of two PSD kernels is PSD, a product of two PSD kernels is PSD, etc.

Question **8**

Complete

Mark 5.00 out of 5.00

Consider the kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \alpha + (1 - \alpha)\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$$

where α is a parameter of the kernel.

Give the conditions on the parameter α for which the kernel k is positive semi-definite.

sum_i{ sum_j{ cicj(alpha+(1-alpha)<x,x'> }}

=alpha sum_i{ sum_j{ cicj}} + (1-alpha) sum_i{ sum_j{ cicjxixj}}

=alpha sum_i {(ci)^2} + (1-alpha) sum_i{(cixi)^2} >= 0

the term inside the sum are both quadratic hence ≥ 0. to ensure that the kernel is positive semi-definite, alpha needs to be between 0 and one such that 0 ≤ a ≤ 1 and 0 ≤ 1-a ≤ 1

Comment:

$$k(\mathbf{x}, \mathbf{x}') = \alpha + (1-\alpha)\langle \mathbf{x}, \mathbf{x}' \rangle$$

Condition of α for k to be PSD

That is

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \, k(\mathbf{x}_i, \mathbf{x}_j) \geq 0.$$

We first write

$$\begin{aligned} 0 &\leq \alpha\Big(\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \Big) + (1-\alpha)\Big(\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \, \langle \mathbf{x}_i, \mathbf{x}_j \rangle \Big) \\ &\leq \alpha\Big(\sum_{i=1}^{n} c_i \Big)^2 + (1-\alpha)\Big(\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \, \langle \mathbf{x}_i, \mathbf{x}_j \rangle \Big) \end{aligned}$$

Thus, $\alpha = [0, 1]$.

Squared Distance in feature space $\phi(\cdot)$ between $\mathbf{x} = (1, 0)$ and $\mathbf{x}' = (0, 1)$

As k satisfies the Mercer's condition (symmetric and PSD), we know that

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$$

To compute the squared distance in the feature space $\phi(\cdot)$, we first

$$\begin{aligned} \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|^2 &= \Big(\phi(\mathbf{x}) - \phi(\mathbf{x}')\Big)^T \Big(\phi(\mathbf{x}) - \phi(\mathbf{x}')\Big) \\ &= \phi(\mathbf{x})^T \phi(\mathbf{x}) - 2\phi(\mathbf{x})\phi(\mathbf{x}') + \phi(\mathbf{x}')^T \phi(\mathbf{x}') \\ &= k(\mathbf{x}, \mathbf{x}) - 2k(\mathbf{x}, \mathbf{x}') + k(\mathbf{x}', \mathbf{x}'). \end{aligned}$$

Evaluating the terms, we get

- $k(\mathbf{x}, \mathbf{x}) = \alpha + (1-\alpha)(1) = 1$
- $k(\mathbf{x}, \mathbf{x}') = \alpha + (1-\alpha)(0) = \alpha$
- $k(\mathbf{x}', \mathbf{x}') = \alpha + (1-\alpha)(1) = 1$

Thus,

$$\begin{aligned} \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|^2 &= 2(1) - 2(\alpha) \\ &= 2(1-\alpha) \end{aligned}$$

Explicit feature map when $d = 1$

$$\phi(x) = \begin{bmatrix} \sqrt{\alpha} \\ \sqrt{(1-\alpha)}\,x \end{bmatrix}$$

Question **9**

Complete

Mark 0.00 out of 5.00

A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that is symmetric and positive semi-definite typically induces a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^h$, that relates to the kernel as $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$.

Consider the kernel defined in the question above, and assume that the number of input dimensions is $d = 2$. *Compute* the square distance in feature space between the point $\boldsymbol{x} = (1, 0)$ and $\boldsymbol{x}' = (0, 1)$, i.e. compute

$$\|\phi((1,0)) - \phi((0,1))\|^2.$$

(Hint: You may rewrite this quantity in terms of kernel computations).

|| alpha + (1-alpha)(1+0) - (alpha + (1-alpha)(0+1)) || = || 0 ||

0

Comment:

Question **10**

Complete

Mark 0.00 out of 5.00

Consider now the case where $d = 1$ (i.e. one-dimensional input data). *Give* an explicit feature map $\phi : \mathbb{R} \to \mathbb{R}^h$ associated to the kernel k.

alpha +(1-alpha)x^2

Comment:

Question **11**

Complete

Mark 3.00 out of 5.00

For the feature map you have found in the question above, *give* a point in the feature space $\mathbb{R}^h$ that has no pre-image in the input space, i.e. for which it is impossible to find a corresponding data point x.

-1

the above mentioned feature map cannot be < 0 since alpha +(1-alpha)x^2 since for every x and alpha where 0 < alpha < 1 the result is >= 0.

Comment:

answer is correct w.r.t to the answer provided for Q10, gives 3 points

Question **12**

Complete

Mark 2.50 out of 5.00

Consider the optimization problem

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x}\|^2 \qquad \text{subject to} \qquad \boldsymbol{w}_1^\top \boldsymbol{x} \geq 1 \quad \text{and} \quad \boldsymbol{w}_2^\top \boldsymbol{x} \leq -1$$

Viewing $\boldsymbol{x} \in \mathbb{R}^d$ as the input and $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$ as two feature detectors, *give* a high-level interpretation of such optimization problem, as well as an interpretation of Slater's conditions for duality.

we want to find a hyperplane where all wi lie outside of the decision boundary. Through the optimization problem we want to maximise the margin. Hyperplane's orientation is characterised by vector x for which the hyperplane is orthogonal. Through the maximisation problem we can find the optimal x.

Due to the problem involving a norm, this task represents a convex type of problem. Thus we can use Slater's condition. If we can satisfy the constraints with strict inequality, then the solution of the optimization problem can be also written by its Lagrange dual formulation.

Comment:

Question **13**

Complete

Mark 2.50 out of 5.00

Rewrite the optimization problem above in the form of

$$\max_{\boldsymbol{\alpha}} \Big\{ \min_{\boldsymbol{x}} \Big\{ \dots \Big\}\Big\}$$

where $\boldsymbol{x}$ and $\boldsymbol{\alpha}$ are the primal and dual optimization variables respectively.

argmax over alpha{argmin over x{0.5*||x||^2 +alpha*(|w'x|-1)}}

Comment:

Question **14**

Complete

Mark 5.00 out of 10.00

Give the dual optimization problem associated to the primal optimization problem above. In particular, *state* the objective and the constraints on the dual optimization variables.

L = 0.5*||x||^2 +alpha(|w'x|-1)

derivative wrt w: x - alpha*w = 0 -> x = alpha*w

set in L: argmax over alpha{ 0.5*|| alpha*w||^2 +alpha(|w*alpha*w|-1) }

objective 0.5*|| alpha*w||^2 constraint |w*alpha*w| >= 1

Comment:

Question **15**

Complete

Mark 5.00 out of 5.00

For this specific problem, *explain* in one or two sentences when (or whether) the dual formulation should be preferred over the primal formulation.

Primal is preferred when N is very large(>> data points). Dual is better to use if the dimension D is very large.

Comment:

Information

Consider a supervised dataset $(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $t_i \in \mathbb{R}$. The input data is stored in a matrix of size $N \times d$, and the corresponding outputs (targets) are stored in a vector of size N.

We consider a kernel-based regression model given by:

$$f(\boldsymbol{x}) = \frac{\sum_{i=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_i) \cdot t_i}{\sum_{i=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_i)}$$

and we would like to implement it.

Your implementation should take the form of a *function* that receives four arguments,

1. a matrix called Xtrain containing the training data points,
2. a corresponding vector of targets called Ttrain,
3. the hyperparameter(s) of the kernel,
4. a matrix of test points you would like to predict called Xtest.

Your function should return the vector of predicted values for points in Xtest. Your implementation should be *efficient*, i.e. make use of numpy/scipy vector and matrix operations when possible, and avoid redundant computations. *(Hint: you may make use of the function scipy.spatial.distance.cdist).*

Question **16**

Complete

Mark 7.00 out of 10.00

Implement the regression model defined above when the kernel function is given by $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma \cdot \|\boldsymbol{x} - \boldsymbol{x}'\|^2)$.

```
getKernel(X1,X2,gamma):
D2 = scipy.spatial.distance.cdist(X1,X2,"sqeuclidean")
K = numpy.exp(-gamma*D2)
return K



def regression(X1,X2,t):

return getKernel(X1,X2,gamma).dot(t)/getKernel(X1,X2,gamma)
```

Comment:

Question **17**

Complete

Mark 0.00 out of 10.00

Write a function that selects the best kernel hyperparameter for the model above. Use for this a holdout validation procedure where the training data is randomly split into 80% training and 20% validation, where candidate hyperparameters are spaced logarithmically between 10^{-5} and 10^{5}, and where the mean square error is used as a selection criterion.

```
Xtrain,Xtest = X[R[:len(R)//1.6]]*1,X[R[len(R)//1.6:]]*1 #80%
Ttrain,Ttest = T[R[:len(R)//0.8]]*1,T[R[len(R)//0.8:]]*1 #20%
```

Comment: