ML Course – Exam question examples

Prof. Dr. Isabel Valera

Disclaimer: *In the following we provide an overview of the course contents that are relevant for the exam. In addition, we provide some examples of the concepts that may be asked in the exam. Note that the questions below are not directly exam questions, which may come in different forms and phrasings, but just examples of the concepts that you should know. They are also not complete, and thus any other concept discussed in a discussed in a lecture and/or corresponding tutorial and pointed out below is material suitable to be tested during the exam.*

Block I

- L3: Everything
  Q1: *Given the posterior probability of a a binary decision task, compute the optimal decision rule? Compute its error probability. Can this error be further reduced?*
  Q2: *Given a regression task, can we obtain the optimal regressor independently of the considered loss function? Assuming the squared loss, what is the form of the Bayes optimal regressor? Does it agree with the Bayes optimal regressor with respect to the L1-loss?*
  Q3: *If the Bayes classifier is optimal, why don't we use it in practice?*

- L4: Everything
  Q1: *Consider a binary classification task. How does the solution of the ERM problem assuming logistic loss relates with the maximum likelihood estimator of the logistic regression classifier?*
  Q2: *Given two classifiers with similar performance in terms of accuracy but one with the double number of parameters than the other. According to the Occam's Razor principle, which of the two classifiers should we pick and why?*

Block II

- L5 & 6: Everything
  Q1: *Explain with your own words the difference between lasso and ridge regression, explaining the pros and cons of each approach.*
  Q2: *What are the main assumptions of the Gauss-Markov theorem? In case they are fulfilled, for a given training dataset with* $N = 100$ *observations, each with* 1000 *features, is the least squared solution the best linear regressor to apply? Why?*

Q3: *Based on the geometric interpretation of Least Squares regression and Ridge regression, how does the predictions made by the two models compare?*

- L7: Everything
  Q1: *Compare LDA and logistic regression classifiers in terms of their formulation and assumptions?*
  Q2: *Given a training dataset with pairs of input features and binary labels, compute the optimal weighting vector for the LDA classifier.*
  Q3: *Given a training dataset with pairs of input features and labels, how would you learn the parameters of a logistic regression classifier? If we add an L1-regularization term to the logistic regression loss, how the solution to this regularized ERM problem relates to the MAP estimator of the parameters?*

- L8 & 9: Everything
  Q1: *Given the predictions of a classifier in a test set, compute accuracy, F1-score, precision and recall.*
  Q2: *Which performance metrics would you use to evaluate the performance of a classifier in an unbalanced classification problem and why?*
  Q3: *Given the ROC curves for several classifiers, which one would you select for deployment and why?*
  Q4: *Given the training, validation and test errors of different classifiers, pick the best model and report its generalization error.*

Block III

- L11 - L14: Functions that preserve convexity, convexity of a learning, dual problem, SVM formulation.
  Q1: *Does the regularized logistic regression with L1-regularization have a uniqe solution? Why?*
  Q2: *Derive the dual problem of a cost-sensitive soft-margin SVM explaining each of the steps.*
  Q3: *Show that* $\exp g(x)$ *is convex if* g *is convex.*
  Q4: *Derive the dual optimization problem for ridge regression detailing each step.*

- L15 & 16: Everything
  Q1: *Derive kernel ridge regression step by step, pointing out at which step the representer theorem and the kernel trick play a role.*
  Q2: *Does kernel least squares regression satisfy the representer theorem? Why (yes/no)?*
  Q3: *Compute the kernel for the mapping function* $\Phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$.
  Q4: *If* k_1 *and* k_2 *are positive definite kernels, is the new kernel* $k(x, x') = f(x)k_1(x, x')k_2(x, x')f(x')$ *a valid positive definite kernel? Why?*

Block IV

- L18 - L19: (Mostly from review of probability theory) Only conditional independence, conditional expectation... Specific knowledge on fair classification will not be required.

Block V

- L20: Everything but EM algorithm derivation.
  Q1: *For a given dataset, compute two steps of the Lloyd's algorithm.*
  Q2: *Compare the kmeans solution with the one of agglomerative hierachical clustering assuming single linkage.*
  Q3: *What are the main assumptions and parameters of the GMM for clustering?*

- L21: PCA, whitening, and ICA idea and comparison with PCA (not the algorithm).
  Q1: *Compute the whitened version of a given dataset.*
  Q2: *What are the key assumptions of PCA? Is it suitable for non-linear data? In the negative case, how can we overcome such a limitation?*
  Q3: *Name a key difference in the assumptions made by PCA and ICA.*
  Q4: *Derive the Kernel PCA formulation explaining each step. How does its solution compare with (standard) PCA?*

Block VI

- L22-25: Key concepts NN design and learning.
  Q1: *Why is it important to regularize the training of NNs? Name an approach to regularize NNs and explain in your own words the main idea behind it.*
  Q2: *Compute the output size of a 2D convolutional layer with input size $W \times H$, filter $K \times K$, and striding equal to s.*
  Q3: *Given the training, validation and test errors of a NN architecture trained with different batch sizes, select the best model and report its generalization error.*