Mid-term exam, Machine Learning (MDS), April 20th, 2023

Instructions:

- You have **1h** to solve the exam
- Please return this paper with your answers, make sure to write your name **clearly**
- Mark whether the following statements are **true** or **false**, or leave blank
- Correct answers count +1 point, incorrect answers count -1, non-answered answers count 0
- At least **half** of the questions must be answered
- The mid-term grade is given by the formula $10 \exp\left(\frac{\text{nr. of correct} - \text{nr. of incorrect questions}}{35} - 1\right)$

(only marked FALSE ones)

General

- Regression and clustering are types of supervised learning **FALSE: clustering is unsupervised**
- Clustering and dimensionality reduction are types of unsupervised learning
- Machine learning is particularly useful when we try to solve a problem that is easy to program however data is scarce **FALSE: if the problem is easy to program, then there is no need for ML**
- Preprocessing is a task that can often be automated **FALSE: preprocessing requires significant time and expertise and trial and error**
- In supervised learning, we attempt to predict a target value from feature values describing an object
- In supervised learning, we always generate models with minimum training error **FALSE: in fact, we would prefer some small trianing error to zero error with an overfitted model, for example**
- Empirical risk, the opposite of training error, serves as an approximation to the true risk **FALSE: empirical risk is the same as training error, so not the opposite**

Bayes and probabilities

- Bayes theorem can be derived from the product rule of probability theory
- Bayes theorem transforms prior distributions into posterior distributions
- $P(Y) = \sum_x P(Y|X = x)P(X = x)$ for X, Y discrete random variables
- $P(Y) = \sum_x P(X = x|Y)P(X = x)$ for X, Y discrete random variables **FALSE**
- Expert information on the domain is encoded into the model through the posterior distribution **FALSE: through the prior distribution**
- The posterior distribution contains both expert information on the domain and information gathered through observation (data)
- The likelihood function is a probability distribution over the possible values of the parameters for a model **FALSE: it is not a probability distribution**

Regression

- Least squares linear regression is obtained by assuming Gaussianity on the input variables **FALSE: Gaussianity is assumed for the error term of the target or output variable**
- Linear regression can produce non-linear predictions if we apply linear transformations on the input variables **FALSE: to produce non-linear predictions, we must apply non-linear transformations**
- The best choice in linear regression is to minimize square error **FALSE: not necessarily, e.g. in the presence of outliers it would be better to minimize *absolute error***
- High bias models will tend to underfit
- Low variance models will tend to overfit **FALSE: high variance models tend to overfit**
- Lasso regression uses a form of regularization that is useful in the presence of outliers **FALSE: Lasso uses L1 norm to penalize complexity, as a result it produces feature selection; for outliers we would need L1 error over the error term in regularized expression**
- The GCV for ridge regression computes the LOOCV error exactly **FALSE: GCV is an approximation to LOOCV**

Model selection, resampling and errors

- Resampling methods are useful to learn a model's parameters **FALSE: to learn model's parameters we use learning algorithms, resampling is for model selection and hyper-parameters**
- Resampling methods are useful to learn a model's hyper-parameters
- Cross-validation is used to estimate generalization error
- Cross-validation is used for model selection
- LOOCV is a type of resampling method that can be used as an alternative to cross-validation **FALSE: LOOCV is not an alternative but a particular case**
- In the presence of scarce data, k-fold cross-validation with high values of k is preferable to low values of k for estimating generalization if possible
- Minimizing validation error is a good methodology to ensure good generalization
- Minimizing training error is a good methodology to ensure good generalization **FALSE: training error is not a good measure to use as it may produce overfitted models**

Clustering

- K-means and EM are both methods for learning Mixture of Gaussian models
- The EM algorithm refines a suboptimal solution obtained by k-means until a global optimum is found **FALSE: EM can get stuck on local optima as well**
- K-means is a particular case of EM for Gaussian Mixtures when covariance matrices are assumed diagonal **FALSE: not diagonal, but isotropic or spherical with standard deviation tending to 0**
- Mixing coefficients for the Gaussian mixture are estimated in EM directly from the best soft assignments obtained so far
- In EM, the log-likelihood cannot decrease after each iteration
- In k-means it is possible to get stuck on a local optimum however EM solves this problem **FALSE: EM can get stuck on local optima as well**