Machine Learning - MIRI Master (Final quiz - June 8, 2015)

Tick **clearly** the claims that you think are **true** - deliver just **this sheet** back

Name:

1. **Complexity control and all that**.

F ❑ If we keep complexity low, we do not need to care about training error

F ❑ Training error is always smaller (or equal) than test error

T ❑ Supplying more training data reduces the chances to obtain an overfitted model

F ❑ Regularization penalizes models that are either simpler or more complex than needed

F ❑ Training error is enough to perform model selection

F ❑ The VC dimension for a two-class classifier penalizes training sample size The VC dimension measures the capacity of a model to classify datasets of various sizes and complexities. It does not directly penalize training sample size

F ❑ The VC dimension for a two-class classifier is the maximum number of linear separations that the classifier can perform The VC dimension is the maximum number of points that can be shattered (perfectly classified) by the classifier, not the number of linear

T ❑ In order to check that the VC dimension is (at least) some integer k, we just need to find k points that can be shattered

T ❑ Checking that the VC dimension is infinite requires an infinite number of checks

T ❑ A two-class classifier with infinite VC dimension must have an infinite (or very large) number of parameters

2. **Bayesian classifiers**.

T ❑ The Bayes formula converts prior distributions into posterior distributions

F ❑ The Bayes formula is of theoretical importance, but can never be used in practice

T ❑ The numerator in Bayes formula is enough to perform classification

T ❑ The Bayes classifier is the best possible classifier when the prior and posterior distributions are known

T ❑ For normally distributed classes, Bayesian classifiers turn out to be quadratic discriminant functions (QDA)

F ❑ For normally distributed classes, statistical independence among all variables yields linear discriminant functions (LDA) LDA assumes equal covariance among classes, not statistical independence among variables.

F ❑ For normally distributed classes, Bayesian classifiers are minimum-distance classifiers Bayesian classifiers minimize classification error using probability distributions, not necessarily distance.

T ❑ The Naive-Bayes classifier assumes statistical independence among all variables

F ❑ The kNN classifier can be explained as a Bayesian classifier The kNN classifier is based on distance metrics, not probabilistic models.

F ❑ The kNN classifier works better with more neighbours, although it is computationally more costly kNN does not necessarily work better with more neighbors; there is an optimal k value that balances bias and variance. More neighbors can lead to higher bias.

3. **Maximum Likelihood and GLMs**.

T ❑ The likelihood of a sample is its density for a given choice of parameters

F ❑ The likelihood of a sample is a function of the sample

F ❑ Logistic regression is a generative linear classifier Logistic regression is a discriminative classifier, not generative.

F ❑ Logistic regression assumes normally distributed classes Logistic regression does not assume normal distribution of classes; it models the probability of class membership directly.

F ❑ In a Generalized Linear Model, the prediction is the logistic function applied to a linear function of the predictors This is specific to logistic regression. In GLMs, different link functions can be used (not necessarily logistic)

F ❑ The solution for Logistic regression can be found analytically by minimizing the log-likelihood The solution for logistic regression is typically found using iterative optimization techniques, not analytically

F ❑ In Poisson regression, we are interested in predicting integer outcomes, which are equally likely Poisson regression predicts count data, but the outcomes are not necessarily equally likely

F ❑ In a Generalized Linear Model, we always find the logistic function in one way or another (so it is called the link function) GLMs can use various link functions, not just the logistic function.

F ❑ In statistics, bias and variance are opposite concepts: increasing one must decrease the other

T ❑ Variance always decreases with increasing sample size; however, bias can increase or stay the same

4. **Regression, linear and non-linear**.

F ❑ The regression function is the best possible model in regression, and achieves zero error on the training data The best possible model (regression function) minimizes the expected prediction error, not necessarily achieving zero training error.

T ❑ The risk is equal to the sum of the (squared) bias, the variance and the noise variance

F ❑ The theoretical MSE does not depend on the regression function The theoretical Mean Squared Error (MSE) depends on the chosen regression function and the underlying data distribution

F ❑ Models that are "more complex than needed" will tend to have a large bias and large variance

F ❑ Models that are "less complex than needed" will tend to have a small bias and small variance

T ❑ A linear combination of non-linear (fixed) functions of the inputs make a linear model

F ❑ Ridge regression adds a penalty term to a linear model such that the new model is non-linear

T ❑ Non-linear functions of the data can be estimated by using linear fitting techniques

T ❑ Both RBFs and MLPs create non-linear models by learning adaptive regressors (regressors with parameters)

T ❑ Regularization allows the specification of models that are more complex than needed; it also helps numerically

5. **Kernels and SVMs**.

F ❑ The kernel function defines kernel matrices whose elements are always positive Kernel matrix elements are not necessarily positive; they can be zero or negative

F ❑ The kernel function defines kernel matrices whose elements are always non-negative

F ❑ The product of two kernel functions may not be a kernel; however, the sum of two kernel functions is always a kernel The sum and product of two valid kernels are both valid kernels.

T ❑ By choosing a valid kernel, we get an inner product in a Hilbert space

F ❑ In order to kernelize a learning algorithm, this must be a supervised one, like the SVM Unsupervised algorithms can also be kernelized, such as kernel PCA.

T ❑ The cost parameter (C) in a SVM acts as a regularizer of the solution

F ❑ Increasing the margin in a SVM leads to larger VC dimension Increasing the margin generally decreases the VC dimension, promoting better generalization.

T ❑ Increasing the margin in a SVM leads to greater chances to separate the data

F ❑ Increasing the value of C in a SVM, the margin of the solution may increase Increasing C usually leads to a smaller margin as the model tries to classify more training points correctly, potentially overfitting.

F ❑ Increasing the value of C in a SVM, the number of training errors may increase Increasing C usually decreases the number of training errors but may increase the risk of overfitting.

6. **Miscellaneous**.

F ❑ The k-means algorithm converges to a global optimum if the number of iterations goes to infinity k-means converges to a local optimum, not necessarily a global one.

T ❑ A Gaussian mixture model assumes normality of the training data

F ❑ The k-means algorithm is used to fine-tune a Gaussian mixture model after the latter has converged k-means is not used to fine-tune GMMs. Instead, the Expectation-Maximization (EM) algorithm is used.

T ❑ The backpropagation algorithm computes the partial derivatives of the error function with respect to the network weights

T ❑ The backpropagation algorithm must be coupled with an optimization method (update rule) to make it a learning algorithm for a MLP

T ❑ A MLP requires the specification of the number of hidden neurons; this is best done by trial-and-error, monitoring the fitting error of the network

F ❑ RBF neural networks are a particular case of MLP networks

F ❑ In a RBF neural network, regularization does not make sense, because it is based on Euclidean distance

F ❑ Feature selection can never increase the practical performance of a learning method, it only reduces learning time

T ❑ Feature selection can be performed after feature extraction, using the extracted features as new variables for selection