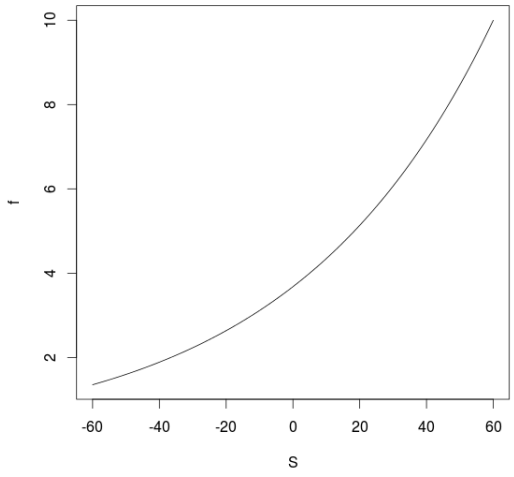
Machine Learning - MIRI Master (Final quiz - June 9, 2016)

Name:

Instructions:

- tick **clearly** the claims that you think are **true** with a $\surd$
- tick **clearly** the claims that you think are **false** with a $\times$
- if you want to "withdraw" an already ticked box, black it out as ■ (it will count now as unanswered)
- all questions are equally weighted (the headings define **blocks** of **ten** questions each)
- there is no obligation to answer individual questions, but at least half (**five**) questions in each block must be answered
- individual question grading: correct answers count $+1$ point, incorrect answers count -1 point; no answer counts 0 points (there are 60 questions = 60 points maximum)
- letting S be the number of points, the overall grade is obtained as

$$f(S) = 10 \exp\left(\frac{S}{60} - 1\right)$$



- deliver just **these sheets** back
- time: 2h+

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

F ❑ Complexity control is necessary in non-linear methods only

T ❑ The empirical error in the training set is always smaller (or equal) than the empirical error in the test set

T ❑ The empirical error in the training set is always smaller (or equal) than the empirical error in the validation set

T ❑ Using a larger validation data set reduces the chances to select an overfitted model

T ❑ Regularization usually penalizes models that are more complex than needed

T ❑ Regularization may penalize models that are simpler than needed

F ❑ Cross-validation guarantees that our model does not overfit the data

Cross-validation helps to estimate the generalization error and can reduce the risk of overfitting by using different subsets of data for training and validation. However, it does not guarantee that the model will not overfit.

F ❑ L_2-regularization produces sparsity, as opposed to L_1-regularization

T ❑ The VC dimension of a two-class classifier is independent of data dimension

F ❑ The VC dimension of a two-class classifier is always a finite integer

2. **Bayesian classifiers.**

T ❑ The Bayes formula transforms prior distributions into posterior distributions

F ❑ The denominator in Bayes formula is enough to perform classification, by taking simply the maximum over the classes

T ❑ The Bayes classifier is the best possible classifier when the prior and class-conditional distributions are known

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

F ❑ For normally distributed classes, equal prior probabilities yield linear discriminant functions

F ❑ The Naive-Bayes classifier can only be used with discrete random variables, because it assumes statistical independence among all variables, given the class

F ❑ The kNN classifier needs no tuning of the number of neighbours, because in the limit of infinite data it is a Bayesian classifier

T ❑ $\sum_b P(a|b)P(b) = 1$, where A, B are discrete random variables

T ❑ $\sum_b P(b|a) = 1$, where A, B are discrete random variables

F ❑ $\sum_b P(a|b) = 1$, where A, B are discrete random variables

3. **Maximum Likelihood and GLMs.**

F ❑ The likelihood is a function of the data sample for a given choice of parameters

F ❑ The negative log-likelihood sometimes yields different results than the likelihood

F ❑ Logistic regression is a linear method that can be used to predict an arbitrary numerical quantity

F ❑ Linear regression assumes normally distributed inputs and outputs

T ❑ In a GLM, the model tries to predict the expected value of the target using a linear function of the predictors and a suitable interface function

Generalized Linear Models (GLMs) use a linear predictor and a link function (interface function) to model the expected value of the target variable.

T ❑ The solution for a GLM can be found analytically by minimizing the log-likelihood, or iteratively using Newton-Raphson

T ❑ In Poisson regression, we are interested in predicting positive outcomes that represent counts

F ❑ The regression function is the best possible predictor, and would achieve zero error on the population

F ❑ In statistics, bias and variance are related concepts: increasing one must increase the other, and viceversa (decreasing one must decrease the other)

F ❑ The mean squared error is always preferred for optimization, because it is the more theoretically sound

4. **Neural networks**.

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

F ❑ A linear combination of non-linear functions with adaptive parameters is a linear model

T ❑ The backpropagation algorithm computes the partial derivatives of the given 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

F ❑ Even if we fix the initial weights, a MLP is a non-deterministic method

T ❑ The activation function for the output neurons is dictated by the nature of the target variable

T ❑ A MLP requires the specification of the number of hidden neurons, which can be done in a variety of ways

F ❑ RBF and MLP neural networks can be seen as a particular case of the same class of neural networks

F ❑ In a RBF neural network there is no regularization, because they are based on Euclidean distances instead of inner products

F ❑ Regularization does not make sense in neural networks, because they learn adaptive regressors (regressors with parameters)

5. **Kernels and SVMs**.

F ❑ The kernel function defines kernel matrices whose elements are always positive

T ❑ Any linear combination of two kernel functions is a kernel function

T ❑ By choosing a valid kernel, we get an inner product in some Hilbert space, which is our new feature space

In SVMs, the Lagrange multipliers (αn) for support vectors are positive and lie between zero and the cost parameter CC. Non-support vectors have αn equal to zero.

F ❑ In SVMs, the Lagrange coefficients α_n are negative for the support vectors only

T ❑ In order to kernelize a learning algorithm, this must be based on Euclidean distances or inner products of the data

Kernelization involves replacing inner products with kernel functions. Algorithms based on Euclidean distances or inner products can be kernelized using this approach.

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

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

T ❑ Increasing the value of C in a SVM, the number of training errors cannot increase

F ❑ The VC dimension of a SVM depends on the data distribution

The VC dimension of an SVM depends on the capacity of the classifier (e.g., the number of support vectors and the choice of kernel) rather than the data distribution itself.

T ❑ The VC dimension of a SVM depends on the margin we allow

6. **Miscellaneous**.

F ❑ The k-means algorithm converges to a global optimum as the number of iterations goes to infinity

T ❑ A Gaussian mixture model assumes that the data has been generated by some finite mixture of Gaussians

T ❑ The k-means algorithm can be used to initialize a Gaussian mixture model

T ❑ A Random Forest is "random" partly because the *variables* used in *each decision tree* are optimized amongst a randomly chosen subset

T ❑ A Random Forest is "random" partly because the *variables* used in *each decision node* are optimized amongst a randomly chosen subset

T ❑ A Random Forest is "random" partly because the *data* used in *each decision tree* come from a different bootstrap resample

F ❑ A Random Forest is "random" partly because the *data* used in *each decision node* come from a different bootstrap resample

F ❑ In Machine Learning, there is no limit on the achievable predictive performance of a model, it is just a matter of choosing the correct method, and tuning the parameters

T ❑ In Machine Learning, pre-processing can make a large impact on learning, and therefore on predictive performance

T ❑ A system (living or not) learns when it uses past experience to improve future performance

---