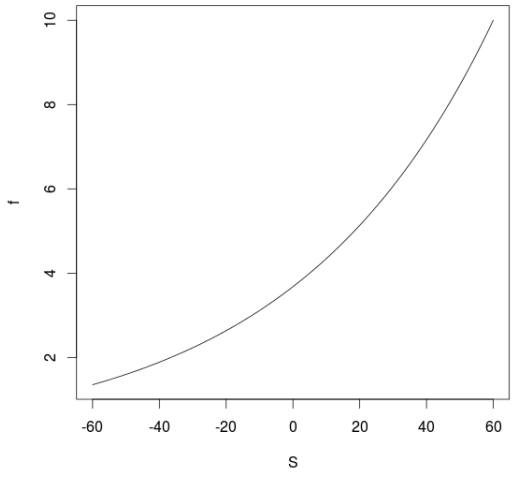
Machine Learning - MIRI Master (Final quiz - June 9, 2016)

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. False
2. False
3. False
4. True
5. True
6. False
7. False
8. False

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

   - ❏ Complexity control is necessary in non-linear methods only
   - ❏ The empirical error in the training set is always smaller (or equal) than the empirical error in the test set
   - ❏ The empirical error in the training set is always smaller (or equal) than the empirical error in the validation set
   - ❏ Using a larger validation data set reduces the chances to select an overfitted model
   - ❏ Regularization usually penalizes models that are more complex than needed
   - ❏ Regularization may penalize models that are simpler than needed
   - ❏ Cross-validation guarantees that our model does not overfit the data
   - ❏ L_2-regularization produces sparsity, as opposed to L_1-regularization
   - ❏ The VC dimension of a two-class classifier is independent of data dimension
   - ❏ The VC dimension of a two-class classifier is always a finite integer

2. **Bayesian classifiers**.

   - ❏ The Bayes formula transforms prior distributions into posterior distributions
   - ❏ The denominator in Bayes formula is enough to perform classification, by taking simply the maximum over the classes
   - ❏ The Bayes classifier is the best possible classifier when the prior and class-conditional distributions are known
   - ❏ For normally distributed classes, Bayesian classifiers turn out to be quadratic discriminant functions
   - ❏ For normally distributed classes, equal prior probabilities yield linear discriminant functions
   - ❏ The Naive-Bayes classifier can only be used with discrete random variables, because it assumes statistical independence among all variables, given the class
   - ❏ The kNN classifier needs no tuning of the number of neighbours, because in the limit of infinite data it is a Bayesian classifier
   - ❏ $\sum_b P(a|b)P(b) = 1$, where A, B are discrete random variables
   - ❏ $\sum_b P(b|a) = 1$, where A, B are discrete random variables
   - ❏ $\sum_b P(a|b) = 1$, where A, B are discrete random variables

1. True
2. False
3. True
4. True
5.
6. False
7. False
8. False
9. True
10. False

3. **Maximum Likelihood and GLMs**.

   - ❏ The likelihood is a function of the data sample for a given choice of parameters
   - ❏ The negative log-likelihood sometimes yields different results than the likelihood
   - ❏ Logistic regression is a linear method that can be used to predict an arbitrary numerical quantity
   - ❏ Linear regression assumes normally distributed inputs and outputs
   - ❏ 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
   - ❏ The solution for a GLM can be found analytically by minimizing the log-likelihood, or iteratively using Newton-Raphson
   - ❏ In Poisson regression, we are interested in predicting positive outcomes that represent counts

1. False
2. False
3, False
4. False
5.
6.
7.
8. False
9. False
10. False

- ❑ The regression function is the best possible predictor, and would achieve zero error on the population
- ❑ In statistics, bias and variance are related concepts: increasing one must increase the other, and viceversa (decreasing one must decrease the other)
- ❑ The mean squared error is always preferred for optimization, because it is the more theoretically sound

4. **Neural networks**.

- ❑ Non-linear functions of the data can be estimated by using linear fitting techniques
- ❑ A linear combination of non-linear functions with adaptive parameters is a linear model
- ❑ The backpropagation algorithm computes the partial derivatives of the given error function with respect to the network weights
- ❑ The backpropagation algorithm must be coupled with an optimization method (update rule) to make it a learning algorithm for a MLP
- ❑ Even if we fix the initial weights, a MLP is a non-deterministic method
- ❑ The activation function for the output neurons is dictated by the nature of the target variable
- ❑ A MLP requires the specification of the number of hidden neurons, which can be done in a variety of ways
- ❑ RBF and MLP neural networks can be seen as a particular case of the same class of neural networks
- ❑ In a RBF neural network there is no regularization, because they are based on Euclidean distances instead of inner products
- ❑ Regularization does not make sense in neural networks, because they learn adaptive regressors (regressors with parameters)

1. True
2. False
3. True
4. True
5. False
6. True - choose function based on target variable
7. True
8. True
9. False
10. False

5. **Kernels and SVMs**.

- ❑ The kernel function defines kernel matrices whose elements are always positive
- ❑ Any linear combination of two kernel functions is a kernel function
- ❑ By choosing a valid kernel, we get an inner product in some Hilbert space, which is our new feature space
- ❑ In SVMs, the Lagrange coefficients α_n are negative for the support vectors only
- ❑ In order to kernelize a learning algorithm, this must be based on Euclidean distances or inner products of the data
- ❑ The cost parameter (C) in a SVM acts as a regularizer of the solution
- ❑ Increasing the margin in a SVM leads to greater chances to separate the data not necessarily
- ❑ Increasing the value of C in a SVM, the number of training errors cannot increase
- ❑ The VC dimension of a SVM depends on the data distribution
- ❑ The VC dimension of a SVM depends on the margin we allow

1. False
2. False
3. True
4. False
5. True
6. True
7. False
8. True
9.
10.

6. **Miscellaneous**.

- ❑ The k-means algorithm converges to a global optimum as the number of iterations goes to infinity
- ❑ A Gaussian mixture model assumes that the data has been generated by some finite mixture of Gaussians
- ❑ The k-means algorithm can be used to initialize a Gaussian mixture model
- ❑ A Random Forest is "random" partly because the *variables* used in *each decision tree* are optimized amongst a randomly chosen subset

1. False
2. True
3.
4. False
5. True
6. True
7. False
8. False
9. True
10. True

- ❏ A Random Forest is “random” partly because the *variables* used in *each decision node* are optimized amongst a randomly chosen subset
- ❏ A Random Forest is “random” partly because the *data* used in *each decision tree* come from a different bootstrap resample
- ❏ A Random Forest is “random” partly because the *data* used in *each decision node* come from a different bootstrap resample
- ❏ 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
- ❏ In Machine Learning, pre-processing can make a large impact on learning, and therefore on predictive performance
- ❏ A system (living or not) learns when it uses past experience to improve future performance

---