	Max Points	Achieved
Q1 General	34	
Q2 Metric Learning	13	
Q3 Bias vs. Variance	23	
Q4 SVMs	29	
Q5 Boosting etc.	21	
Total	120	

Question 1. General Questions

Circle the correct answers for each question (multiple correct answers possible):

1. [2] The error of a squared loss classifier decomposes into which terms:
    a. Malt, Hops, Yeast
    b. Frustration, Sadness, Hopelessness
    c. Bias, Arrogance and Tradeoff
    d. Noise, Variance and Bias
2. [5] Do you accept the 5 bonus points for filling in the evaluation survey?
    a. Yes, I do.
    b. No, I don't need charity points!
    c. I am so confident, you can subtract 5 points from my score.
3. [2] If you knew the underlying data distribution, which classifier would you use:
    a. k-nearest neighbors
    b. Bayes rule
    c. SVM
    d. Logistic Regression
    e. AdaBoost
4. [4] Which statements are true?
    a. The Perceptron algorithm **always** converges after a **finite** number of steps.
    b. The Perceptron algorithm **always** converges after a **finite** number of steps, if a separating hyper-plane exists.
    c. There **exist linearly separable data** sets, for which the Perceptron algorithm **never converges**.
    d. There **exist linearly separable data** sets, for which the Perceptron algorithm **converges** only in the limit as the number of iterations approaches infinity.

5. [4] What statement about MIRA (passive aggressive update) are correct?

    a. MIRA is identical to the perceptron update

    b. When run over the entire training data set, it converges to exactly the same solution as a linear SVM would.

    c. The MIRA update performs the smallest change in an l2-sense to the weight vector, such that the current input is classified correctly.

    d. The MIRA update performs the smallest change in an l1-sense to the weight vector, such that the current input is classified correctly

6. [7] Which of the following algorithms are non-parametric?

    a. Perceptron

    b. K-Nearest Neighbor Classification

    c. LARS

    d. Decision Trees (full depth, without pruning or early stopping)

    e. Linear SVM

    f. SVM with RBF Kernel

    g. Random Forests

7. [10] For each of the following algorithms formulate the **loss function** and **regularization term**. (If multiple choices are possible, one is sufficient).

    a. SVM:

    b. LASSO:

    c. Ridge-Regression:

    d. Ordinary Least Squares:

    e. Logistic regression:

8. [2] Name at least one reason why someone might prefer l1-regularization to l2-regularization.

9. [2] Name at least one reason why someone might prefer l2-regularization to l1-regularization.

Question 2. Metric Learning

1. [2] Which of the following statements are true?

- **a.** LMNN learns a linear transformation of the data set
- **b.** LMNN learns a large margin separating hyper-plane
- **c.** One part of the LMNN objective is a surrogate loss for the leave-one-out k-NN classification error
- **d.** The LMNN optimization problem is non-convex

2. [4] Which of the following statements are true?

- **a.** PCA is unsupervised
- **b.** The PCA projection minimizes reconstruction error
- **c.** The PCA projection maximizes variance
- **d.** PCA only projects out noise but not any of the signal.

3. [3] You have a correct implementation of LMNN. What are good ideas?

- **a.** To avoid local minima, you initialize with the solution of NCA.
- **b.** To speed up convergence, you set the step-size (learning rate) extremely large.
- **c.** You perform early stopping on a validation data set, and re-train with the same number of iterations on the union of training data and validation data before deployment.

4. [4] Assume your data X is centered. You run [U,S,V]=svd(X); The embedding/projection matrix of which algorithms can you obtain trivially without any further computation (except simple rescaling, division, square-rooting or transposing).

    a. PCA

    b. MDS

    c. Whitening

    d. NCA

Question 4. Support Vector Machines

1. [5] Which of the following statements about SVMs are true.

    a. The SVM optimization problem is convex.

    b. The kernel trick allows SVMs to learn non-linear decision boundaries.

    c. SVM (especially with non-linear kernels) are ideally suited to be boosted

    d. The SVM optimization problem is a linear program.

    e. Support Vectors lie exactly on the decision boundary

2. [8] State the SVM classification rule h(x) for an input x with label y for both the kernel (with kernel function k(,)) and the primal case. Label all variables that you introduce.

3. [8] Formulate what it means to find a maximum margin hyper-plane (in binary classification). Derive how the SVM optimization achieves this through a norm minimization. State clearly what assumptions you are making.

4. [8] Franz von Hahn has a multi-class data set with C classes. He downloads SVM code for binary classification and trains C "1 vs. all" classifiers. During test-time he applies a test input to all classifiers and returns the label with the highest response. He is shocked that although each "1 vs. all" classifier has a reasonable train and validation error on its own, his combined classifier performs rather poorly. What would you recommend him to change about his setup? Provide a detailed description of what he needs to do.

Question 5. Boosting / p>>n / l1-regularization

1. [7] Which of the following statements about Adaboost are true.

   a. For some classifiers it reduces the training error at a logarithmic rate.

   b. For some classifiers it reduces the training error at an exponential rate.

   c. For some classifiers it reduces bias.

   d. It is typically very robust against label noise.

   e. It learns a linear decision boundary.

   f. It can always reduce the training error to zero, without any further assumptions on the data or classifier.

   g. Adaboost finds the optimal step-size with a line search in each iteration.

2. [6] Which of the following statements about LARS are true?

   a. Typically after m iterations you have selected m non-zero variables.

   b. After every LARS iteration, there exists a Lasso problem for which the current weight vector is an optimal solution.

   c. For every Lasso problem, there exists an integer m such that after m iterations of LARS (for some m>=0) the weight vector is an optimal solution to the optimization problem.

   d. LARS reduces the contribution to the gradient of all dimensions in the active set exponentially.

   e. LARS updates all weights of all dimensions in the active set by the same constant in each iteration.

   f. LARS requires the user to set the step-size carefully.

3. [8] Write down a pseudo-code implementation of gradient boosted regression trees (GBRT). (You can assume that you have a working implementation of CART.)