UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
Department of Electrical and Computer Engineering

ECE 544NA Pattern Recognition

Midterm Exam

Fall 2013

Mark Hasegawa-Johnson Thursday, November 21, 2013

Problem 1

[40 pts] Consider a binary classification problem where $p(x|y=1)$ is uniformly distributed over the rectangle described by the vertices $V^{(1)} = \{(-2,0),(2,0),(-2,2),(2,2)\}$ and $p(x|y=-1)$ is uniformly distributed over the triangle described by the vertices $V^{(-1)} = \{(-2,0),(2,0),(0,4)\}$. We want to minimize the 0-1 loss (probability of error) in the following scenarios [Hint: geometry is your friend]

(a) Assume a uniform prior: $p(y=1) = p(y=-1) = \frac{1}{2}$. Find $f_B(x)$, the Bayes optimal classifier under the 0-1 loss. Is $f_B(x)$ unique? What is its risk (probability of error)? [15 pts]

(b) Assume a uniform prior. What is the optimal 1-layer neural network (*linear* classifier) under the 0-1 loss? Is it Bayes optimal? [10 pts]

(c) If $p(y = 1) < \frac{1}{2}$, is there a 1-layer network that achieves Bayes optimum? How about a 2-layer network (i.e. with *one* hidden layer)? Justify your claims. [15 pts]

Problem 2

[30 pts] A kernel $k(x, y)$ is called a Mercer kernel if it satisfies the Mercer conditions: 1) symmetry, 2) continuity, and 3) positive semi-definiteness. Let $k(x, y)$ be a Mercer kernel and $|k(x, y)| < 1$. Prove each of the following statements if it is true, and disprove/provide a counterexample otherwise.

(a) $\frac{1}{k(x,y)}$ is a Mercer kernel. [5 pts]

(b) $\frac{1}{1+k(x,y)}$ is a Mercer kernel. [10 pts]

(c) $1 - k(x, y)$ is a Mercer kernel. [5 pts]

(d) $\frac{1}{1-k(x,y)}$ is a Mercer kernel. [10 pts]

Problem 3

[30 pts] Please evaluate each of the following statements as $true$ or $false$. Unless otherwise specified, assume the most general case. You **do not** need to justify.

(a) If the function to be optimized is convex, gradient descent converges regardless of the step size.

(b) Maximum a posteriori (MAP) estimation is equivalent to maximum likelihood (ML) estimation when the prior is uniformly distributed.

(c) Linear discriminant analysis (LDA) as a dimensionality reduction tool is always better than PCA when the final goal is classification.

(d) 1-nearest neighbor classification has a lower risk than 3-nearest neighbor classification.

(e) The perceptron does not penalize correct examples.

(f) The perceptron algorithm does not converge when the training dataset is linearly separable.

(g) It is possible to approximate any function to arbitrary precision with a 3-layer neural network that has finitely many nodes.

(h) The Hessian (second derivative) is always invertible because it is symmetric.

(i) A restricted Boltzmann machine (RBM) is a more efficient implementation of PCA when the number of desired dimensions is small.

(j) Two sets are linearly separable if and only if the intersection of their convex hulls is a line.