Practice problems to prepare for Exam 1

COMS 4771 Fall 2016

Problem 1 (Linear classifiers)**.** Which of the following classifiers are *linear classifiers* in $\mathbb{R}^d$?

- The function $f\colon \mathbb{R}^d \to \{0,1\}$ given by $f(\boldsymbol{x}) = \mathbb{1}\left\{\frac{1}{1+\exp(-\langle \boldsymbol{w}, \boldsymbol{x}\rangle)} - \frac{1}{2} > 0\right\}$ for some vector $\boldsymbol{w} \in \mathbb{R}^d$.
- The 1-NN classifier based on the following training data ($d = 2$):

	Feature 1	Feature 2	Label
Example 1	-1	-1	-1
Example 2	$+1$	$+1$	$+1$
Example 3	-1	$+3$	-1
Example 4	$+1$	$+5$	$+1$

- A classifier trained using Kernelized Online Perceptron with the kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) := \langle \boldsymbol{x}, \boldsymbol{A}\tilde{\boldsymbol{x}}\rangle$, where $\boldsymbol{A}$ is a $d \times d$ symmetric positive definite matrix.

Problem 2 (Maximum likelihood estimation)**.**

(a) Consider the model $\mathcal{P}$ of distributions over the positive integers $\mathbb{N}$ with probability mass functions given by

$$P_\theta(x) = (1-\theta)^{x-1}\theta, \quad x \in \mathbb{N}$$

for parameter $\theta \in (0,1)$ called the *success parameter*.

Derive a formula for the maximum likelihood estimator for the success parameter given data $\{x_i\}_{i=1}^n$ (treated as an iid sample).

(b) Consider the model $\mathcal{P}$ of probability distributions over the non-negative reals $\mathbb{R}_+$ with probability densities on the given by

$$p_\lambda(x) = \lambda e^{-\lambda x}, \quad x \in \mathbb{R}_+$$

for parameter $\lambda > 0$ called the *rate parameter*.

Derive a formula for the maximum likelihood estimator for the rate parameter given data $\{x_i\}_{i=1}^n$ (treated as an iid sample).

Problem 3 (Kernels)**.** Let $K_1\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $K_2\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be kernel functions.

(a) Define $K_3\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by

$$K_3(\boldsymbol{x}, \boldsymbol{x}') \;:=\; K_1(\boldsymbol{x}, \boldsymbol{x}') + K_2(\boldsymbol{x}, \boldsymbol{x}')\,.$$

Is K_3 always a kernel function (whenever K_1 and K_2 are)? Give a brief justification for your answer.

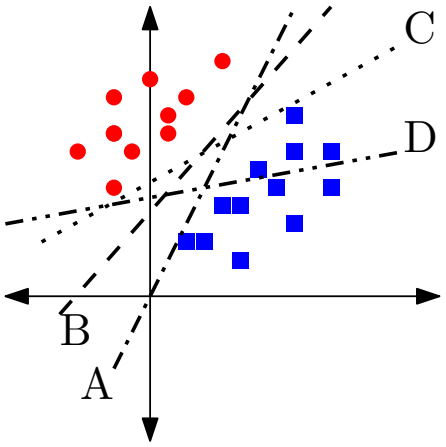
(b) Define $K_4\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by

$$K_4(\boldsymbol{x}, \boldsymbol{x}') \;:=\; -K_1(\boldsymbol{x}, \boldsymbol{x}') - K_2(\boldsymbol{x}, \boldsymbol{x}')\,.$$

Is K_4 always a kernel function (whenever K_1 and K_2 are)? Give a brief justification for your answer.

(c) Let S be a data set of n labeled examples from $\mathcal{X} \times \{-1, 1\}$. Furthermore, let $\boldsymbol{\phi} \colon \mathcal{X} \to \{\boldsymbol{z} \in \mathbb{R}^D : \|\boldsymbol{z}\|_2 = 1\}$ be the feature mapping corresponding to the kernel $K \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Suppose the shortest weight vector $\boldsymbol{w} \in \mathbb{R}^D$ such that $y\langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle \geq 1$ for all $(\boldsymbol{x}, y) \in S$ has length $\|\boldsymbol{w}\|_2 = 10$. At most how many times does Kernelized Online Perceptron compute the value of the kernel function K when run on input S? Give as small of an upper bound as possible, and briefly justify your answer.

Problem 4 (Linear classifiers, again)**.** The figure below depicts the decision boundaries of four linear classifiers (labeled A, B, C, D), and the locations of some labeled training data (positive points are squares, negative points are circles).



Consider the following algorithms for learning linear classifiers which could have produced the depicted classifiers given the depicted training data:

1. (Batch) Perceptron
2. Online Perceptron (making one pass over the training data)
3. ERM for homogeneous linear classifiers
4. An algorithm that exactly solves the SVM problem

What is the most likely correspondence between these algorithms and the depicted linear classifiers? Give a brief explaination for your matching.

Problem 5 (True/false)**.** Determine if each of the following statements is true or false. (For your own edification, you should also come up with a brief justification for your answer.)

(a) For any classifier $f \colon \mathcal{X} \to \{0, 1\}$, if S is an iid sample from a distribution P over $\mathcal{X} \times \{0, 1\}$, then $\mathrm{err}_S(f)$ is an unbiased estimator of $P(f(X) \neq Y)$, where $(X, Y) \sim P$.

(b) It is possible for a leaf in a decision tree (during greedy training) to have zero "classification error" but non-zero "Gini index".

(c) Suppose an iid sample from a distribution P over $\mathcal{X} \times \{0, 1\}$ is randomly partitioned into three sets, S, V, and T. Further, suppose a feature map $\boldsymbol{\phi} \colon \mathcal{X} \to \mathbb{R}^D$ is determined based on $S \cup V$, and a classifier $\hat{f} \colon \mathbb{R}^D \to \{0, 1\}$ is trained using $\{(\boldsymbol{\phi}(\boldsymbol{x}), y) : (\boldsymbol{x}, y) \in S\}$. Then $\mathrm{err}_V(\hat{f} \circ \boldsymbol{\phi})$ is an unbiased estimator of the true error rate of f.

(d) Consider the setting from (c) through the penultimate sentence. Then $\mathrm{err}_T(\hat{f} \circ \boldsymbol{\phi})$ is an unbiased estimator of the true error rate of f.

(e) A classifier based on a generative model with Gaussian class conditional distributions is a linear classifier if the class conditional distributions are *product distributions*.

(f) The linear classifier with the smallest training error rate on a set of examples S can be obtained as the solution to a linear program.

(g) If Perceptron is run on a linearly separable data set S, and $\boldsymbol{w}_\star$ is the shortest vector satisfying $y\langle \boldsymbol{w}_\star, \boldsymbol{x}\rangle \geq 1$ for all $(\boldsymbol{x}, y) \in S$, then the angle between weight vector $\hat{\boldsymbol{w}}_t$ maintained by Perceptron and $\boldsymbol{w}_\star$ decreases after every update.

(h) Every kernel function $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ satisfies $K(\boldsymbol{x}, \boldsymbol{x}') \geq 0$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$.

(i) Suppose $(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n)$ is the solution for the SVM dual problem when the (linearly separable) training data is $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Let $S' := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n+m}$ be a superset of S, and let $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{n+m})$ be the solution for the SVM dual problem when the training data is S'. For each $i = 1, 2, \ldots, n$, whenever $\hat{\alpha}_i = 0$, then $\hat{\beta}_i = 0$. That is, if $(\boldsymbol{x}_i, y_i)$ is not a support vector for the SVM classifier with data set S, then it is not a support vector for the SVM classifier with data set S'.

(j) The soft-margin SVM optimization problem has no solution whenever the data set S contains a pair of examples $(\boldsymbol{x}, y)$ and $(\boldsymbol{x}, y')$ where $y \neq y'$.