Problem 1 (General Machine Learning) **3 + 3 + 3 + 5 = 14 Points**

a) Which of the following are advantages of neural networks? Check all that apply.

☐ The training procedure scales well with large amounts of data.

☐ The training procedure is a convex optimization problem.

☐ They perform well on some classically difficult machine learning problems.

b) A function $f : \mathbb{R} \to \mathbb{R}$ is convex if...

☐ f'' is strictly increasing.

☐ f' is always non-negative

☐ f'' is strictly positive.

c) Write down the names of two regression algorithms.

d) Qualitatively describe the difference between the hard and soft-margin support vector machine.

Problem 2 (Support Vector Machines) **3 + 4 + 5 + 4 = 16 Points**

Let $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$, $i = 1, \dots, n$, be classification training data. Consider the following variation on the soft margin support vector machine:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{2n}\sum_{i=1}^{n} \xi_i^2 \\ \text{s.t.} \quad & y_i\left(\langle w, x_i\rangle + b\right) \geq 1 - \xi_i \text{ for } i = 1, 2, \dots, n \\ & \xi_i \geq 0 \text{ for } i = 1, 2, \dots, n. \end{aligned}$$

Note that the slack parameter penalization is now squared.

a) Explain why we can drop the constraints $\xi_i \geq 0$.

b) Construct the Lagrangian.

c) Take the derivative of the Lagrangian wrt w, b, and ξ and use the result to derive the dual problem.

d) Once we have found the optimal dual variables, how would we find the optimal w, b?

Problem 3 (Kernels) **3 + 3 + 4 + 5 = 15 Points**

a) Describe a situation where it would be better not to kernelize an algorithm.

b) Describe a situation where a polynomial kernel would be an optimal choice of kernel.

c) Construct a kernel different from the kernels defined in class and demonstrate why it is a kernel. You may use kernels defined in class and kernel properties described in class to help you construct it.

d) Let k and k' be kernels on $\mathbb{R}^d$ and suppose we know that there exists functions $\phi : \mathbb{R}^d \to \mathbb{R}^D$ and $\phi' : \mathbb{R}^d \to \mathbb{R}^D$ such that $k(x, y) = \phi(x)^T \phi(y)$ and $k'(x, y) = \phi'(x)^T \phi'(y)$ for all $x, y \in \mathbb{R}^d$. Prove that g is a kernel where $g(x, y) = k(x, y)k'(x, y)$.

Problem 4 (Regression) **3 + 3 + 4 + 5 = 15 Points**

a) Recall that the loss function for lasso regression is

$$\min_w \frac{1}{2}\lambda \|w\|_1 + \left\|y - X^T w\right\|^2,$$

where $\|w\|_1 = \sum_{i=1}^d |w_i|$. What is an advantage of lasso regression over ridge regression?

b) Can lasso regression be kernelized? Argue why or why not (you do not need to prove this formally).

c) There does not exist a closed form solution for w in lasso regression so one must use gradient descent to optimize it. Evaluate the gradient of the lasso loss function at a point w. You may assume no entry of w is 0.

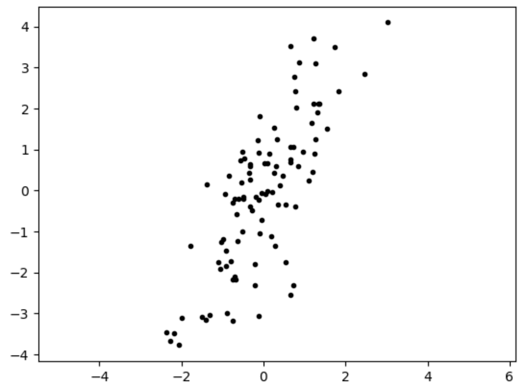
d) Linear regression with offset has a loss function

$$\arg\min_{w,b} \sum_{i=1}^n \left(y_i - \left(w^T x_i + b\right)\right)^2.$$

Find a closed form solution for the minimizer of this loss function. You may use tricks mentioned in the homework solutions to simplify the loss function.

Problem 5 (Principal Component Analysis) **3 + 5 = 8 Points**

a) The following is a scatter plot of some data. Draw the approximate direction of the first principal component on this plot.



b) Given a symmetric matrix S, any unit vector v which satisfies

$$v^T S v = \max_{w:\|w\|=1} w^T S w.$$

is an eigenvector of S associated with the largest eigenvalue of S.

The first principal component, p, of a collection of centered data $x_1, \ldots, x_n$ satisfies

$$p = \arg \max_{u:\|u\|=1} \sum_{i=1}^{n} \left(u^T x_i\right)^2 .$$

Show that this condition is equivalent to p being the eigenvector associated with the largest eigenvalue of the covariance matrix.

Problem 6 (k-Means Algorithm) **6 + 3 = 9 Points**

a) Suppose we are running k-means on a collection of data $x_1, \ldots, x_n$. Given centers $c_1, \ldots, c_m$, write pseudocode describing *one* update of centers using the k-means algorithm.

b) How do we know when to stop running the k-means algorithm?

Problem 7 (Neural Networks) **$3 + 3 + 3 + 5 = 14$ Points**

a) The cross-entropy loss is a commonly used loss function when training neural networks for. . .

   ☐ regression.

   ☐ classification.

b) Write down the equation for, describe, or draw a graph of the sigmoid function.

c) Write down the name of a method for regularizing neural networks.

d) Describe how a convolutional layer in a neural network works in 5 sentences or less.

Problem 8 (k-Nearest Neighbors Classification) **3 + 6 = 9 Points**

a) In a binary classification setting, why is the k-nearest neighbors classification algorithm a bit simpler when k is chosen to be odd?

b) Let $(x_1, y_1), \ldots, (x_n, y_n)$ be some classification data. Given a test point x_t, write pseudocode for predicting the test label y_t using the k-nearest neighbors algorithm.
