Exercises
Introduction to Machine Learning
FS 2018

Series 3, Mar 20th, 2018
(Perceptron, Feature Selection, Kernels)

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s18`
Email questions to:
Stephanie Hyland, hyland@inf.ethz.ch

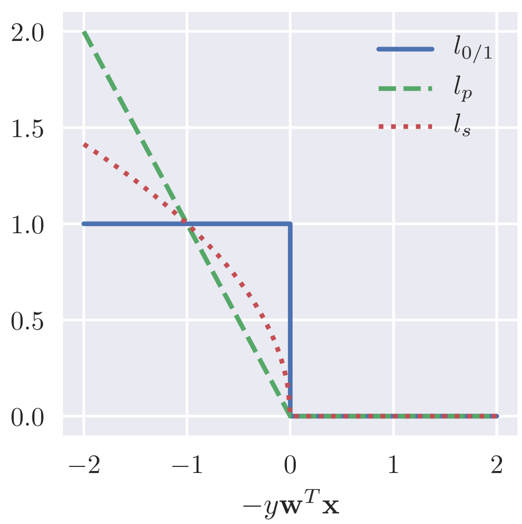
We will publish sample solutions on Friday, Mar 30th.

Problem 1 (Perceptron/SVM):

1. a) How does the perceptron algorithm relate to stochastic gradient descent?
   b) How does the perceptron objective relate to the support vector machine objective?
   c) Write down the training objective for the SVM and derive the gradient updates using stochastic gradient descent. Assume a minibatch size of B.

2. The perceptron in its original formulation uses a 0/1 loss function (shown below, solid). A surrogate loss function $l_p(\mathbf{w};\mathbf{x},y) = \max(0, -y\mathbf{w}^T\mathbf{x})$ is instead used in optimisation (dashed). We see that this surrogate loss is a poor match for the 0/1 loss near zero. Suppose we try (shown in dotted line):

$$l_s(\mathbf{w};\mathbf{x},y) = \begin{cases} 0, & \text{for } \operatorname{sign}(\mathbf{w}^T\mathbf{x}) = y \\ \sqrt{-y\mathbf{w}^T\mathbf{x}}, & \text{for } \operatorname{sign}(\mathbf{w}^T\mathbf{x}) \neq y \end{cases}$$



   a) Show that $f(x) = \sqrt{x}$ is not convex.
   b) Show that $f(x) = x^p$ is convex for all $p \in \mathbb{N}_{>0}$ and $x \in [0, \infty)$. (Hint: use properties of derivatives of convex functions.)

Problem 2 (Feature Selection):

(Exercise 13.5 from Machine Learning: A Probabilistic Perspective by Kevin P. Murphy) We covered ridge (l_2-regularised) and l_1-regularised (lasso) regression in class. A hybrid version called *elastic net* exists which uses both l_1 and l_2 regularisation terms:

$$J_{\mathsf{EL}} = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|^2$$

Defining

$$J_2 = \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{w}\|^2 + c\lambda_1 \|\mathbf{w}\|_1$$

where $c = (1 + \lambda_2)^{-1/2}$ and

$$\tilde{\mathbf{X}} = c \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_d \end{pmatrix}, \qquad \tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_{d\times 1} \end{pmatrix}$$

show that

$$\text{arg min}_{\mathbf{w}} J_{\mathsf{EL}}(\mathbf{w}) = c(\text{argmin}_{\mathbf{w}} J_2(\mathbf{w})))$$

This implies that an elastic net problem can be solved as a lasso problem, using modified data.

Problem 3 (Kernels):

a) For $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, and $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + 1)^2$, find a feature map $\phi(\mathbf{x})$, such that $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}')$.

b) For the dataset $X = \{\mathbf{x}_i\}_{i=1,2} = \{(-3, 4), (1, 0)\}$ and the feature map $\phi(\mathbf{x}) = [x^{(1)}, x^{(2)}, \|\mathbf{x}\|]$, calculate the **Gram matrix** (for a vector $\mathbf{x} \in \mathbb{R}^2$ we denote by $x^{(1)}, x^{(2)}$ its components).