Introduction to machine learning 89-511, Fall 2022: Home Assignment 4

Gal Chechik

Submission date: 2022-12-19

Submit Instructions

Submit to the 'submit' system one file:
- 'solutions.pdf' with your solutions to all questions.

1. (20 pts) Logistic regression

(a) Prove that the logistic regression classifier is equivalent to a softmax over a linear multiclass classifier for two classes $y =$ "a", $y =$ "b", where their separating hyperplanes obey $\mathbf{w}_a = -\mathbf{w}_b$.

(b) Consider a softmax function for k classes that is scaled by a constant $T \in R$

$$f_i(\mathbf{x}) = \frac{\exp(\frac{1}{T}\mathbf{w_i}^T\mathbf{x})}{\sum_k \exp(\frac{1}{T}\mathbf{w_k}^T\mathbf{x})} \tag{1}$$

Show that when $T \to 0$, the softmax converges to the max function. What happens when $T \to \infty$?

(c) Write the gradient update rule for a logisitic regression model, when the usual loss of the negative log likelihood is now regularized with the square of the L_2 norm over the weight vector $\frac{1}{2}||\mathbf{w}||^2$.

2. (10 pts) Support Vector Machines

At class we discussed the SVM formulation

$$\min_{\mathbf{w},\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i \tag{2}$$

where C is a trade-off hyper parameter that weighs the cost of misclassification o(thrugh the slack variable). You are given a binary classification problem for which the cost of misclassifying positive samples is different than the cost of misclassifying negative examples. Formulate the problem as an SVM with two types of slack variables. How many variables are now optimized over in your optimization problem, and what are their dimensions?

3. (10 pts) Polynomial kernels

Given a pair of samples in input space $\mathbf{x}, \mathbf{z} \in R^2$, show that the polynomial kernel of degree 3, $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}^T \mathbf{z})^3$ can be expressed as a dot product in feature space $\phi(\mathbf{x})^T \phi(\mathbf{z})$ with the right choice of a feature map $\phi : R^2 \rightarrow R^d$ $(d > 2)$. Write the explicit form of ϕ.

How many operations are needed to compute $K(\mathbf{x}, \mathbf{z})$, and how many are needed for computing $\phi(\mathbf{x})\phi(\mathbf{z})$? How does the number of operation changes for polynomial kernel of higher degrees?