STAT 535 Homework 6
Out November 16, 2022
Due November 30, 2022
©Marina Meilă
mmp@stat.washington.edu

Problem 1 – SVM solution

a. Show that the value of b is theoretically the same no matter what particular support vector is used for its calculation. You can assume linear SVM and linearly separable data for simplicity, or prove the general case. Hint: take 2 different support vectors x^1, x^2 and show that they can't give you different b values.

b. Assume now that you have a non-linear SVM, defined by a kernel $k(\ ,\ )$ and by the feature map $\phi : \mathbb{R}^n \to \mathcal{H}$. In this case, $w \in \mathcal{H}$, hence it cannot be represented explicitly. However, its norm in $\mathcal{H}$ can be computed as $\|w\| =< w, w >$. Give a simple expression for $\langle w, w \rangle$ that can be computed using the kernel. Assume that you have solved the dual problem and that the dual variables $\alpha_{1:N}$, as well as b are known.

Problem 2 – Leave one out CV and support vectors

This problem is ahead of the CV lecture, but it contains all the information needed for you to solve it. Assume the data set $\mathcal{D}$ contains n samples. You perform *leave-one-out cross-validation* i.e, for $i = 1 : n$ you compute a linear support vector machine classifier f_{-i} on $n-1$ points, leaving out (x^i, y^i). More precisely, f_{-i} is a SVM trained on $\mathcal{D}_{-i} = \mathcal{D} \setminus \{(x^i, y^i)\}$.

a. Assume that the original data set is linearly separable. Prove that each of the n support vector problems is also linearly separable.

b. Is it possible that $f_{-i}(x) \equiv f_{-j}(x)$ for $i \neq j$ two points in the training set $\mathcal{D}$? Give a short motivation or proof.

c. Denote by $\hat{L}_{01}^{loo}$ the error rate in leave-one-out CV, i.e

$$\hat{L}_{01}^{loo} \ = \ \frac{|\{i,\, f_{-i}(x^i) \neq y^i\}|}{n}$$

Prove that $\hat{L}_{01}^{loo} \leq \frac{\#\text{support vectors of } f}{n}$, where f is the linear support vector classifier trained on all the data.

[Problem 3 – Quadratic kernel – NOT GRADED]

In this problem, the points lie on the real line, there are two classes and we use the polynomial degree 2 kernel $K(x, x') = (1 + xx')^2$.

1. What is the mapping $\phi : R \longrightarrow R^d$ satisfying

$$K(x, x') \ = \ \phi(x)^T \phi(x')$$

   and what is its dimension d?

2. Let the data be $\mathcal{D} = \{ (-1, +1), (0, -1), (1, +1) \}$.

   Compute $\phi(x_i)$ and the Gram matrix for this dataset.

3. Write the expression of the primal SVM problem for this data set. Be specific, give numerical values.

4. Write the expression of the dual SVM problem for this data set. Be specific, give numerical values.

5. This dual problem is small enough that it can be solved "manually". [Hint: you can notice that due to symmetry, $\alpha_1 = \alpha_3$ and turn it into a 2 variable problem.] Show that the solution is $\alpha_1 = \alpha_3 = 1,\ \alpha_2 = 2$.

6. What are the values of w and b? Write the expression of the discriminant function $f(x) = w^T\phi(x) + b$. Write the same function now using the kernel K. What are the decision regions of this classifier?

   Make a sketch of the data and the decision regions.

Problem 4 – Online linear regression by Stochastic gradient
This is Problem 3 from Homework 5

Consider the linear regression problem with Least Square loss

$$\min_{\beta} E[(y - \beta^T x)^2] \ = \ \min_{\beta} \ L_{LS} \tag{1}$$

where $y \in \mathbb{R}$, $x \in \mathbb{R}^n$, $\beta \in \mathbb{R}^n$. For simplicity we consider the infinite sample version of the problem, but if you want a variation (ungraded) try also the finite sample version, where we optimize $\hat{L}_{LS}$ instead.

The function in (1) is a quadratic function that has a closed form solution, but we will pretend that we don't know this and investigate the use of (stochastic) gradient descent for this problem.

a. Find the expression of the gradient and Hessian of this problem, i.e $\nabla L_{LS}(\beta)$, $\nabla^2 L_{LS}(\beta)$. Express the Hessian as a function of some well known statistical descriptor(s) of the data distribution.

b. Assume that the covariates x are sampled from a Normal distribution with mean 0 and nonsingular covariance Σ (known). Describe and motivate a reasonable way to find the λ parameter of the Stochastic Gradient algorithm based on this assumption.

c. Write the expression of $d = \frac{\partial L_{LS}(y, \beta^T x)}{\partial \beta}$. Show that the direction of descent d is along x, i.e. $d = \alpha x$ for some scalar α, not necessarily positive. What does the scaling of x represent from a statistical modeling point of view?

e. Write the Stochastic Gradient Descent algorithm to optimize this problem. Assume that λ is known.

For practice, ungraded Repeat the problem with an added regularization term $\frac{C}{2}||\beta||^2$.