Machine Learning and Computational Statistics, Spring 2015
Homework 7: Tikhonov, Ivanov, Square Hinge, and Conditional Density Estimation

Due: Tuesday April 21, 2015, at 4pm (Submit via NYU Classes)

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. You may include your code inline or submit it as a separate file. You may either scan hand-written work or, preferably, write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or MathJax via iPython).

1 Ivanov and Tikhonov Regularization

In lecture there was a claim that the Ivanov and Tikhonov forms of ridge and lasso regression are equivalent. We will now prove a more general result.

1.1 Tikhonov optimal implies Ivanov optimal

Let $\phi : \mathcal{F} \to \mathbf{R}$ be any performance measure of $f \in \mathcal{F}$, and let $\Omega : \mathcal{F} \to \mathbf{R}$ be any complexity measure. For example, for ridge regression over the linear hypothesis space $\mathcal{F} = \left\{ f_w(x) = w^T x \mid w \in \mathbf{R}^d \right\}$, we would have $\phi(f_w) = \frac{1}{n} \sum_{i=1}^n \left(w^T x_i - y_i \right)^2$ and $\Omega(f_w) = w^T w$.

1. (3 pts) Suppose that for some $\lambda > 0$ we have the Tikhonov regularization solution

$$f_* = \arg\min_{f \in \mathcal{F}} \left[\phi(f) + \lambda \Omega(f) \right]. \tag{1}$$

Show that f_* is also an Ivanov solution. That is, $\exists r > 0$ such that

$$f_* = \arg\min_{f \in \mathcal{F}} \phi(f) \text{ subject to } \Omega(f) \leq r. \tag{2}$$

(Hint: Start by figuring out what r should be. If you're stuck on this, ask for help. Then one approach is proof by contradiction: suppose f^* is not the optimum in (2) and show that contradicts the fact that f^* solves (1).)

1.2 Ivanov optimal implies Tikhonov optimal

For the converse, we will restrict our hypothesis space to a parametric set. That is,

$$\mathcal{F} = \left\{ f_w(x) : \mathcal{X} \to \mathbf{R} \mid w \in \mathbf{R}^d \right\}.$$

So we will now write ϕ and Ω as functions of $w \in \mathbf{R}^d$.

Let w^* be a solution to the following Ivanov optimization problem:

$$\begin{aligned} \text{minimize} \quad & \phi(w) \\ \text{subject to} \quad & \Omega(w) \leq r. \end{aligned}$$

Assume that strong duality holds for this optimization problem and that the dual solution is attained. Then we will show that there exists a $\lambda \geq 0$ such that $w_* = \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda\Omega(w)\right]$.

1. (1 pt) Write the Lagrangian $L(w, \lambda)$ for the Ivanov optimization problem.

2. (2 pts) Write the dual optimization problem in terms of the dual objective function $g(\lambda)$, and give an expression for $g(\lambda)$. [Writing $g(\lambda)$ as an optimization problem is expected - don't try to solve it.]

3. (4 pts) We assumed that the dual solution is attained, so let $\lambda^* = \arg\max_{\lambda \geq 0} g(\lambda)$. We also assumed strong duality, which implies $\phi(w^*) = g(\lambda^*)$. Show that the minimum in the expression for $g(\lambda^*)$ is attained at w^*. [Hint: You can use the same approach we used when we derived that strong duality implies complementary slackness [1].] **Conclude the proof** by showing that for the choice of $\lambda = \lambda^*$, we have $w_* = \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda\Omega(w)\right]$.

4. (2 pts, **Optional**) The conclusion of the previous problem allows $\lambda = 0$, which means we're not actually regularizing at all. To ensure we get a proper Ivanov regularization problem, we need an additional assumption. The one below is taken from [1]:

$$\inf_{w \in \mathbf{R}^d} \phi(w) < \inf_{\substack{w \in \mathbf{R}^d \\ \Omega(w) \leq r}} \phi(w)$$

   Note that this is a rather intuitive condition: it is simply saying that we can fit the training data better [strictly better] if we don't use any regularization. With this additional condition, show that $w_* = \arg\min_{w \in \mathbf{R}^d} \left[\phi(w) + \lambda\Omega(w)\right]$ for some $\lambda > 0$.

1.3 Ivanov implies Tikhonov for Ridge Regression.

To show that Ivanov implies Tikhonov for the ridge regression problem (square loss with ℓ_2 regularization), we need to demonstrate strong duality and that the dual optimum is attained. Both of these things are implied by Slater's constraint qualifications.

1. (4 pts) Show that the Ivanov form of ridge regression is a convex optimization problem with a strictly feasible point.

2 Square Hinge Loss and Huberized Square Hinge Loss

The squared hinge loss is a margin loss given by

$$\ell(m) = \left[(1 - m)_+\right]^2,$$

where $(m)_+ = m1(m > 0)$ is the "positive part" of m.

[1] See `https://davidrosenberg.github.io/ml2015/docs/3b.convex-optimization.pdf` slide 24.

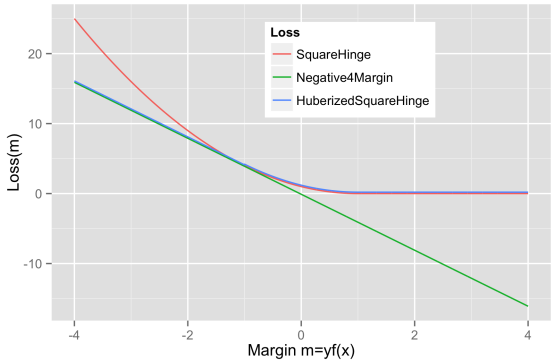

Figure 1: Some margin losses.

1. (2 pts) Suppose we have a training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathcal{X} = \mathbf{R}^d$ and $y_i \in \mathcal{Y} = \{-1, 1\}$, for all $i = 1, \ldots, n$. Consider the linear hypothesis space $\mathcal{F} = \left\{f(x) = w^T x \mid w \in \mathbf{R}^d\right\}$. Write the objective function $J(w)$ for ℓ_2-regularized empirical risk minimization with the square hinge loss over the space $\mathcal{F}$, where $\mathcal{F}$ is parameterized by w.

2. (2 pts) It turns out that $J(w)$ is differentiable at every w. Give the derivative of $J(w)$.

3. (3 pts) Give pseudocode or otherwise explain how you would use stochastic gradient descent to find $w^* = \arg\min_w J(w)$. You need to specify your approach to the step size, but you do not have to specify a stopping criterion, though you may if you like.

4. (2 pts) Assuming that we start SGD at $w = 0$ (or any w in the span of the data), justify the claim that the output of SGD can be written in the form:

$$w = \sum_{i=1}^{n} \beta_i x_i.$$

5. (2 pts) In relation to the SGD algorithm, how would you characterize the x_i's that are and are not support vectors?

6. (2 pts) Show that $J(w)$ is convex. You may use any of the standard facts about convex functions.

7. (2 pts) The "Huberized" square hinge loss (shown in Figure 1) is a margin loss given by

$$\ell(m) = \begin{cases} -4m & m < -1 \\ \left[(1-m)_+\right]^2 & \text{otherwise.} \end{cases}$$

When might you prefer the Huberized square hinge loss to the square hinge loss?

3 Conditional Exponential Distributions

Suppose we want to model the amount of time one will have to wait for a taxi pickup based on the location and the time. The exponential distribution is a natural candidate for this situation. The exponential distribution is a continuous distribution supported on $[0, \infty)$. The set of all exponential probability density functions is given by

$$\text{ExpDists} = \left\{ p_\lambda(y) = \lambda e^{-\lambda y} 1(y \in [0, \infty)) \mid \lambda \in (0, \infty) \right\}.$$

Recall that a family is a **natural exponential family** of continuous distributions on $\mathbf{R}$ with parameter $\theta \in \mathbf{R}$ if its densities can be written as

$$p_\theta(y) = \frac{1}{Z(\theta)} h(y) \exp\left[\theta y\right],$$

where $Z(\theta) = \int h(y) \exp\left[\theta y\right] \, dy$ is the **partition function**. θ is called the **natural parameter**, and the **natural parameter space** Θ consists of all θ for which $Z(\theta) < \infty$. $h(y)$ is called the **base measure.**

1. (4 pts) Write the family of exponential distributions as a natural exponential family. Give expressions for the base measure and the partition function. Identify the natural parameter space.

2. (3 pts) Let $x \in \mathbf{R}^d$ represent the input features from which we want to predict an exponential distribution. We will use a generalized linear model (GLM) approach. Suggest a reasonable function ψ to map $w^T x$ to a value in the natural parameter space Θ. Then write an expression for $p_w(y \mid x)$, the predicted probability density function conditioned on x.

3. (3 pts) Suppose we have a data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathbf{R}^d$ and $y_i \in [0, \infty)$ for $i = 1, \ldots, n$. Give the optimization problem you would solve to fit the GLM we have been discussing to training data $\mathcal{D}$.

4. (5 pts, **Optional**) Suppose we think that a linear function of x doesn't extract enough information, and we'd like to use a more expressive model. For full credit, explain how you would use gradient boosting in this situation. For partial credit, present another reasonable approach to this problem.

References

[1] Marius Kloft, Ulf Brefeld, Sören Sonnenburg, Pavel Laskov, Klaus-Robert Müller, and Alexander Zien. Efficient and accurate lp-norm multiple kernel learning. *Advances in neural information processing systems*, 22(22):997–1005, 2009. 4