Homework 5: Loss functions

Due: Thursday 11/18/2021 9:15am

1. *Loss functions*
   In our first look at regression in this course, we showed how to predict $y \in \mathbf{R}$ given $x \in \mathbf{R}^d$ by finding a vector w minimizing the least squares loss

   $$\|y - Xw\|^2.$$

   This problem is also called ℓ_2 regression, and the loss is called a quadratic loss. However, now that we have grown more sophisticated both in modeling and in optimization, we understand that the quadratic loss is not always the best choice, and that it can be beneficial to use regularization to ensure model interpretability or to improve generalization.

   Please list at least two cases where we should use a loss function that is not quadratic. For each, state the input space $\mathcal{X}$, the output space $\mathcal{Y}$, describe the loss function and regularizer you would use for this problem (and, optionally, any feature transformations), and explain why your choice of loss function and regularizer make sense for this problem. Feel free to use a problem you've encountered in your class project.

2. *Quantile Regression.*

   You can find data and starter code for this problem in the Jupyter notebook `quantileRegression.ipynb`, available at

   `http://github.com/orie4741/homework/blob/master/quantileRegression.ipynb`

   We will be using a random sample of datapoints taken from the 2015 Natality Data of the National Center for Health Statistics. We are interested in investigating the effect of gender, mother's marital status, and prenatal care in the first trimester on the baby's birth weight.

   (a) Fit an ordinary least squares regression to the data. Interpret the coefficients that you find.

(b) Fit a quantile regression on the data for the 5th quantile $q = 0.05$ and for the 95th quantile $q = 0.95$. What do these models predict, and how does it differ from the prediction of the least squares regression? Compare these coefficients to those you found in part a).

(c) Fit quantile regressions for $q = 0.05, 0.10, \ldots, 0.95$.

(d) Create an intercept plot that plots quantiles against the intercept coefficient from that quantile regression. Create coefficient plots for MaritalStatus, Male, and PrenatalCare coefficients. How do the coefficients change as the quantile increases?

(e) What is the meaning of the intercepts of the quantile regressions?

(f) What does the coefficient plot tell you about the effect of prenatal care for infants with low birth weight compared to those with average birth weights?

Hint: you may find that the solution has very large (even infinite!) norm. You may want to add small quadratic regularization to stabilize the answer. If you use a regularizer, you'll likely want to normalize your loss by the number of examples n.

3. *Multiclass classification and ordinal regression.* In this problem, we will study some important properties of loss functions for multiclass classification and ordinal regression.

(a) In class we have defined the multinomial logit function as follows. To predict nominal data with k classes, identify each class with an integer $1, \ldots, k$, and let $W \in \mathbf{R}^{k \times d}$, $x \in \mathbf{R}^d$, so the prediction vector $z = Wx \in \mathbf{R}^k$. Define

$$\mathbb{P}(y = i|z) = \frac{\exp(z_i)}{\sum_{j=1}^{k} \exp(z_j)}.$$

(See page 37 of the loss function slides for details.) Define the imputed region for class i as

$$\mathcal{A}_i = \{x : \mathbb{P}(y = i|Wx) \geq \mathbb{P}(y = j|Wx), \forall j \in \mathcal{Y}\}.$$

Explain what the imputed region represents, and show that each imputed region $\mathcal{A}_i$ is convex.

As a reminder, a set S is convex if for any $x \in S$, $x' \in S$, and $0 \leq \lambda \leq 1$,

$$\lambda x + (1 - \lambda)x' \in S.$$

(b) *One-vs-all classification.* In the one-vs-all classification scheme, we define a loss function as

$$\ell(y, z) = \sum_{i=1}^{k} \ell^{\text{bin}}(\psi(y)_i, z_i),$$

where

$$\psi(y) = (-1, \ldots, \overbrace{1}^{y\text{th entry}}, \ldots, -1) \in \{-1, 1\}^k.$$

Here we will use logistic loss as our binary loss function

$$\ell^{\text{bin}}(\psi_i, z_i) = \ell_{\text{logistic}}(\psi_i, z_i) = \log(1 + \exp(-\psi_i z_i)).$$

(See the loss function slides on multiclass classification for details.)

Prove the following inequality and explain what it means:

$$\ell(i, \psi(i)) \leq \ell(j, \psi(i)), \forall i, j \in \mathcal{Y}.$$

Hint: notice the left hand side asks you to essentially "plug ψ twice", since

$$\ell(i, \psi(i)) = \sum_{j=1}^{k} \ell^{\text{bin}}(\psi(i)_j, \psi(i)_j).$$

(c) *Ordinal regression.* One method for ordinal regression is to define a loss function

$$\ell(y, z) = \sum_{i=1}^{k-1} \ell^{\text{bin}}(\psi(y)_i, z_i),$$

where

$$\psi(y) = 2(\mathbb{1}(y > 1), \mathbb{1}(y > 2), \ldots, \mathbb{1}(y > k-1)) - 1 \in \{-1, 1\}^{k-1}.$$

Again, we will use logistic loss as our binary loss function ℓ^{bin}. (See page 42 of the loss function slides for details.)

Prove the following inequalities hold, and explain what they mean:

$$\ell(i, \psi(i)) \leq \ell(j, \psi(i)), \forall i, j \in \mathcal{Y}.$$

$$\ell(i+1, \psi(i)) \leq \ell(i+2, \psi(i)), \forall i \in \mathcal{Y}.$$

4. *Hinge loss vs. logistic loss.* In class we defined hinge loss

$$\ell_{\text{hinge}}(x, y; w) = (1 - y w^T x)_+$$

and logistic loss

$$\ell_{\text{logistic}}(x, y; w) = \log(1 + \exp\left(-y w^T x\right)).$$

Suppose we want to minimize the regularized empirical risk

$$\min \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; w) + \lambda \|w\|_2^2,$$

where λ. In this problem, we see how each of these loss functions performs on a binary classification problem.

The problem is to predict if a breast tumor is benign or malignant based on its features. The dataset, `breast-cancer.csv`, can be found at https://github.com/ORIE4741/homework/blob/master/breast-cancer.csv

The dataset consists of 683 data points. The first column is the class (-1: benign, 1: malignant), and the following 9 columns are the features.

Starter code for this problem that specifies parameter values can be found in the Jupyter notebook
`hinge_loss_logistic_loss.ipynb` available at

http://github.com/orie4741/homework/blob/master/hinge_loss_logistic_loss.ipynb

(a) Split the data set randomly into training (50%), validation (25%), and test (25%) sets.

(b) Using validation to select the regularization parameter λ, fit a linear support vector classifier to minimize the regularized empirical risk with a hinge loss. Do the same for a logisitic regressor.

(c) Remember the misclassification rate is defined as

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(\hat{y}_i \neq y_i),$$

where $\hat{y}_i$ is your prediction for test data point i, and $\mathbb{1}(\hat{y}_i \neq y_i)$ is 1 when $\hat{y}_i \neq y_i$ and 0 otherwise.

Report the misclassification rates of w^*_{hinge} and w^*_{logistic} on the test set. Which model performs better?

(d) Logistic loss can be interpreted as the negative log likelihood of y given $w^T x$

$$\ell_{\text{logistic}}(x, y; w) = -\log \mathbb{P}^{\text{logistic}}(x, y; w),$$

so

$$\exp\left(-\ell_{\text{logistic}}(x, y; w)\right) = \mathbb{P}^{\text{logistic}}(x, y; w).$$

Similarly, we can give hinge loss a probabilistic interpretation:

$$\frac{1}{z(x; w)}\exp\left(-\ell_{\text{hinge}}(x, y; w)\right) = \mathbb{P}^{\text{hinge}}(x, y; w),$$

where

$$z(x; w) = \exp\left(-\ell_{\text{hinge}}(x, 1; w)\right) + \exp\left(-\ell_{\text{hinge}}(x, -1; w)\right)$$

is the normalizing constant. Why is there no normalizing constant for logistic loss?

(e) Compute the log likelihood of these two models

$$\sum_{i=1}^{n} \log(\mathbb{P}^{\text{logistic}}(x_i, y_i; w_{\text{logistic}}))$$

and

$$\sum_{i=1}^{n} \log(\mathbb{P}^{\text{hinge}}(x_i, y_i; w_{\text{hinge}}))$$

using the test data set and report the log likelihood. Which one is larger?

5. *Bias and Variance in Tree-Based Bagging Ensemble Methods.*

   Fill out the Jupyter notebook `BiasVarianceTree.ipynb`, available at

   `http://github.com/orie4741/homework/blob/master/BiasVarianceTree.ipynb`