Machine Learning

V. Adamchik CSCI 567 Spring 2019

Discussion 4 University of Southern California

Gradient Descent
Perceptron
Logistic Regression

Problem 1

Why is the Hessian of logistic loss positive semidefinite.

Solution. By definition for any u: $u^T H u \geq 0$

We know that for the logistic loss
$H = \sigma(1-\sigma)\, x^T x$, where σ is the sigmoid function.

Compute $u^T H u$

$$u^T H u = \sum_{n=1}^{N} u^T \sigma(w^T x_n)(1-\sigma(w^T x_n))\, x_n^T x_n u$$

$$u^T H u = \sum_{n=1}^{N} \sigma(w^T x_n)(1-\sigma(w^T x_n))(u^T x_n)^2 \geq 0$$

since $0 < \sigma < 1$.

Problem 2

Can we apply Newton's method to the perceptron loss to minimize classification error?

$$F(\boldsymbol{w}) = \sum_{n=1}^{N} \max(0, -y_n \boldsymbol{w}^T x_n)$$

Solution

Apply Newton's method to perceptron.

$$F(w) = \sum_{n=1}^{N} \max(0, -y_n w^T x_n) \qquad x_{n+1} = x_n - H^{-1}(x_n)\nabla f(x_n)$$

Compute the gradient:

$$\nabla F(w) = -\sum_{n=1}^{N} y_n x_n \; I(mistake\ on\ x_n)$$

Compute the Hessian: $H(w) = 0$

Problem 3

Which of the following surrogate losses is not an upper bound of the 0-1 loss?

(A) perceptron loss $\max\{0, -z\}$
(B) hinge loss $\max\{0, 1-z\}$
(C) logistic loss $\log(1 + \exp(-z))$
(D) exponential loss $\exp(-z)$

Solution: A

Problem 4

The following table shows a binary classification training set and the number of times each point is misclassified during a run of the perceptron algorithm. What is the final output of the algorithm? Assume $w^{(0)} = 0$.

$\boldsymbol{x}$	y	Times misclassified
(-3, 2)	+1	5
(-1, 1)	-1	5
(5, 2)	+1	3
(2, 2)	-1	4
(1, -2)	+1	3

Solution: (0, -3)

Problem 5

Suppose we obtain a hyperplane w via logistic regression and are going to make a randomized prediction on the label y of a new point x based on the sigmoid model. What is the probability of predicting y = +1?

$(a)\ \boldsymbol{e}^{-\boldsymbol{w}^T\boldsymbol{x}}$

$(b)\ \dfrac{1}{1+\boldsymbol{e}^{-\boldsymbol{w}^T\boldsymbol{x}}}$

$(c)\ \dfrac{1}{1+\boldsymbol{e}^{\boldsymbol{w}^T\boldsymbol{x}}}$

$(d)\ \mathbb{I}[\boldsymbol{w}^T\boldsymbol{x} \geq 0]$

Solution: b

Problem 6

Assume we have a training set ($\mathbf{x}_1$, y_1), …, ($\mathbf{x}_N$, y_N), the probability of seeing out come y is given by

$$P(\boldsymbol{y}|\boldsymbol{x}_n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{w}^T\boldsymbol{x}_n)^2}{2\,\sigma^2})\right)$$

Find the maximum likelihood estimations for $\boldsymbol{w}$ and σ

Solution

The probability of seeing the outcomes y_1 ,…, y_N is

$$P(\boldsymbol{w}) = \prod_{\boldsymbol{n=1}}^{N} P(y_n | \boldsymbol{x}_n)$$

Taking the negative log, this becomes

$$F(\boldsymbol{w}) = N \ln \sqrt{2\,\pi} + N \ln \sigma + \frac{1}{2\,\sigma^2} \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^T \boldsymbol{x}_n\right)^2$$

Maximizing P is the same as minimizing F, which is the same objective as for linear regression. Therefore,

$$\boldsymbol{w}_* = \left(X^T X\right)^{-1} X^T \boldsymbol{y}$$

Solution

Next we minimize F(**w**) with respect to σ

by setting the derivative to zero

$$\frac{N}{\sigma} - \frac{1}{\sigma^3} \|X\, \boldsymbol{w}_* - \boldsymbol{y}\|_2^2 = 0$$

Solving for σ gives the MLE estimate

$$\sigma = \frac{1}{\sqrt{N}} \|X\, \boldsymbol{w}_* - \boldsymbol{y}\|_2 = \frac{1}{\sqrt{N}} \left\| X\left(X^T X\right)^{-1} X^T \boldsymbol{y} - \boldsymbol{y} \right\|_2$$