The Mathematics of Machine Learning
Homework Set 7

Due 28 April 2022 before 13:00
via Canvas

You are allowed to work on this homework in pairs. One person per pair submits the answers via Canvas. Make sure to put both names on the submission.

1 Understanding Support Vector Machines

For both questions, consider binary classification with $Y \in \{-1, +1\}$ and $X \in \mathbb{R}^d$. For any $\beta \in \mathbb{R}^d$, $\beta_0 \in \mathbb{R}$, the corresponding linear classifier classifies a new input X according to $\hat{Y} = \text{sign}(f_{\beta,\beta_0}(X))$, where

$$f_{\beta,\beta_0}(X) = X^\top \beta + \beta_0.$$

The hinge loss (used in SVMs) and logistic loss (used in logistic regression) are defined as follows:

$$\begin{aligned} L_{\text{hinge}}(Y, f(X)) &= \max\{0, 1 - Yf(X)\} && \text{(hinge loss)} \\ L_{\text{logistic}}(Y, f(X)) &= \ln\left(1 + e^{-Yf(X)}\right) && \text{(logistic loss).} \end{aligned}$$

1. [2 pt] Show that, for any linear classifier f_{β,β_0} and point x^*, the distance of x^* to the decision boundary $\mathcal{B} = \{x \mid f_{\beta,\beta_0}(x) = 0\}$ is $\frac{|f_{\beta,\beta_0}(x^*)|}{\|\beta\|}$.
   Hint: There are different ways to prove this. One approach is as follows:

   (a) *Show that β is a normal vector to the decision boundary. That is, for any $a, b \in \mathcal{B}$, the vector β is orthogonal to the vector $(b - a)$, i.e. $(b - a)^\top \beta = 0$.*

   (b) *Let x' be the projection of x^* onto $\mathcal{B}$. Then, since β is a normal vector of $\mathcal{B}$ (and $\mathcal{B}$ is a $d-1$-dimensional hyperplane), we must have*

   $$x' = x^* + \alpha \frac{\beta}{\|\beta\|},$$

   where $|\alpha|$ is the distance of x^* to $\mathcal{B}$. Plug this into $f_{\beta,\beta_0}(x') = 0$ (because $x' \in \mathcal{B}$) and solve in terms of α.

2. [2 pt] Minimizing some (surrogate) loss function using empirical risk minimization can be viewed as an attempt to approximate the Bayes-optimal predictor for that loss function. We can therefore get more insight into the difference between the hinge loss and the logistic loss by comparing the corresponding Bayes optimal classifiers.

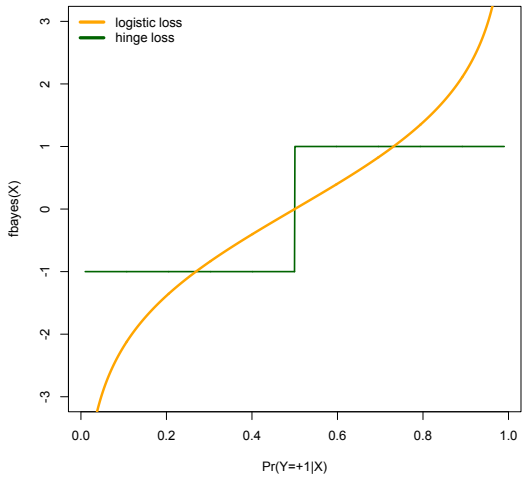

Figure 1: Bayes-optimal classifiers for hinge loss and logistic loss

(a) Show that the Bayes optimal classifier for hinge loss is

$$f_{\text{Bayes}}(X) = \text{sign}\Big(\Pr(Y = +1 \mid X) - 1/2\Big).$$

Hint 1: Abbreviate $p = \Pr(Y = +1 \mid X)$ *to shorten your notation.*
Hint 2: Note that

$$\mathop{\mathbb{E}}_{Y}[\max\{0, 1 - Y\hat{Y}\}|X] = p\max\{0, 1 - \hat{Y}\} + (1 - p)\max\{0, 1 + \hat{Y}\}.$$

(b) Show that the Bayes optimal classifier for logistic loss is

$$f_{\text{Bayes}}(X) = \ln\Big(\frac{\Pr(Y = +1 \mid X)}{\Pr(Y = -1 \mid X)}\Big).$$

See Figure 1 for a plot of the two Bayes optimal functions. Note that the signs of both Bayes optimal classifiers agree, so they will both lead to the same classifications. The difference is that the Bayes optimal decision for the hinge loss is exactly equal to the Bayes optimal decision for the 0/1-loss (corresponding to the idea that SVMs try to estimate the optimal

decision boundary directly), whereas the values for logistic loss provide a smooth interpolation between positive and negative values, which does not have a jump at $\Pr(Y = +1 \mid X) = 1/2$.