November 2, 2017 Lecture 13, Oct. 18, 2017

Kernel Methods

Lecturer: Justin Domke *Scribe: Boya Ren*

1 Multi-class Classification

Before we start kernel methods, we first discuss why we usually have K outputs for a $K-class$ classification instead of $K-1$ outputs.

To formally present the question, suppose we have class label $y \in \{1, 2, \cdots, K\}$. For linear regression, before we formulate it as $f(x) = B^T X$ with $B \in R^{p\times K}$. In the special case of binary classification, i.e., $y \in \{-1, 1\}$, we only need $C \in R^p$. Then for multi-class, what about alternatively setting $B \in R^{p\times(K-1)}$ and "clamp" $f_K(x) = 0$?

The interpretation of using only $K-1$ regressions is: if $f_k(x)$ is high, $f(x)$ "likes" $y = k$, while if all $f_k(s)$ are negative, $f(x)$ "likes" $y = K$. To learn such a model, we need the loss function to be the expression as below.

$$L(y, f(x)) = \begin{cases} -f_y(x) + \log(1 + \sum_{k=1}^{K-1} \exp f_k(x)), \ y \in \{1, 2, \cdots, K-1\} \\ 0 + \log(1 + \sum_{k=1}^{K-1} \exp f_k(x)), \ y = K \end{cases}$$

Compared to the above "reduced output" method, redundancy is good because it is symmetrical and easy to implement. That is why usually we have K outputs for multi-class classification. The comparison of 0-1, hinge and logistic losses is shown in **Fig. 1.1**

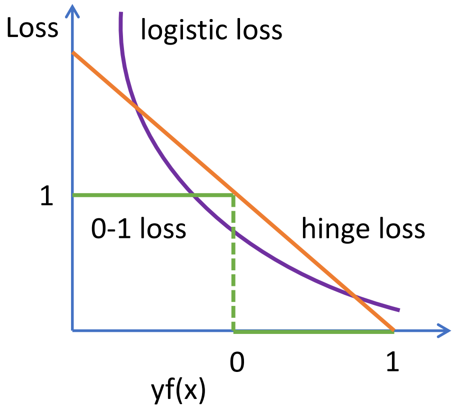

Figure 1.1: Curves of different losses

2 Kernel Methods

Reading: 5.1, 5.2 (skim), 5,8, 12.1, 12.2, 12.3.

2.1 Summary

First it should be noted that “kernel method” does not equal to kernel smoothing, which is used for making the prediction curve smooth, like that of KNN. The major point of kernel methods is combining inner product and locality, which can be summarized below.

- Linear methods are based on inner product (that’s why there are transposes all the time).
- KNN is based on neighborhoods.
- Kernel methods unify these two.
- Learning kernel methods is the major problem.

2.2 Basis Expansion

For a linear predictor $f(x) = \sum_{i=1^p \beta_i x_i}$, the capacity is relatively small compared to non-linear models. To change the capacity of a linear model, here cames basis expansion.

1. Choose $h_m(x): \; R^p \to R$, where $m \in \{1, 2, \cdots, M\}$
2. Fit $f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$

Here are several options for $h_m(x)$

- $h_m(x) = x_m, \; M = P$. This is the original linear model.
- $h_m(x) = x_j^2$ or $x_j x_k$
- $h_m(x) = \log x_j$ or $\sqrt{x_j}$
- $h_m(x) = I[L_m \leq x_k \leq U_m]$

Let’s look at some examples of basis expansion when $p = 1$.

Example 1. Suppose $h_1(x) = 1$, $h_2(x) = x$ and $h_3(x) = x^2$. Then $f(x) = \beta_1 h_1(x) + \beta_2 h_2(x) + \beta_3 h_3(x)$ maybe as shown in the figure below, together with all the basis functions.

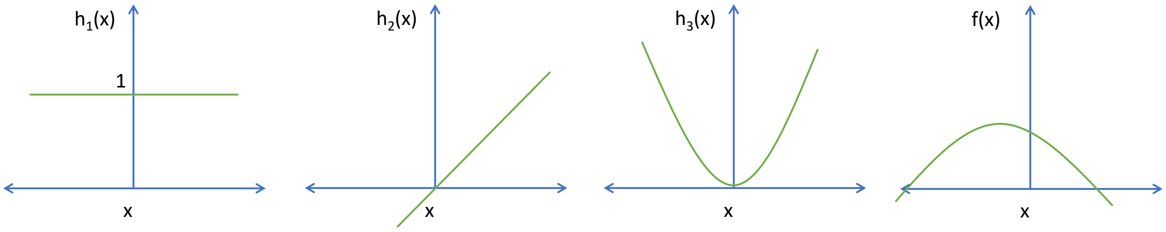

Figure 2.1: Possible Illustration of Example 1

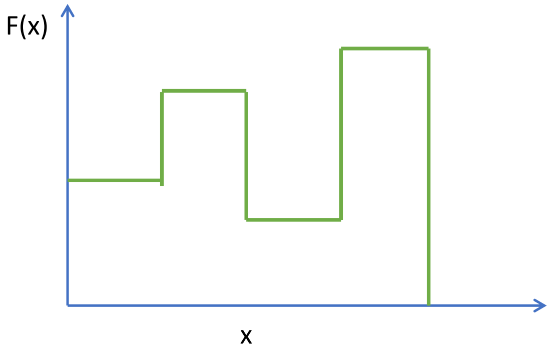

Figure 2.2: Possible Illustration of Example 2

Example 2. Suppose $h_1(x) = I[0 \leq x < 1]$, $h_2(x) = I[1 \leq x < 2]$, $h_3(x) = I[2 \leq x < 3]$ and $h_4(x) = I[3 \leq x \leq 4]$. A possible $f(s)$ is shown is **Fig. 2.2**

There are several tips to choose basis functions:

- Can use domain knowledge
- Neural network: $f(x)V\sigma(WX)$, $h_m(x) = \sigma((WX)_m)$
- Can use basis expansion to reduce capacity, i.e. by dropping features.

Now let's consider the following scenario. We have $p = 100$ and want $h_m(x) = x_{j_1}x_{j_1}\cdots x_{j_d}$ for all combinations of $j_1, j_2, \cdots, j_d \in \{1, 2, \cdots, p\}$. This approach potentially has the problem of overfitting, since $M \approx p^d$.

2.3 Kernel Ridge Regression

First, let's recall the normal linear regression and ridge regression. For linear regression, we have the following relationships.

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N}(y_i - \beta^T x_i)^2 \\ &= (Y - X\beta)^T(Y - X\beta) \end{aligned}$$

$$\nabla RSS(\beta) = -2X(Y - X\beta) = 0$$

$$X^T X\beta = X^T Y$$

$$\beta = (X^T X)^{-1}X^T Y$$

Similarly for ridge regression, we have

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N}(y_i - \beta^T x_i)^2 + \lambda\|\beta\|_2^2 \\ &= (Y - X\beta)^T(Y - X\beta) + \lambda\|\beta\|_2^2 \end{aligned}$$

$$\nabla RSS(\beta) = -2X(Y - X\beta) + 2\lambda\beta = 0$$

$$\beta = (X^T X + \lambda I)^{-1} X^T Y$$

Next time we will continue on the topic of kernel ridge regression.