Predictive Learning

Lecturer: Qiang Sun
January, 2026

Lectures 02

---

1 Supervised Learning

The primary goal of supervised learning is predictive analysis. Predictive analysis refers to a set of techniques that analyze current and historical data to make predictions about the future unknown observation. Within supervised learning, there are two main types of techniques we will consider: regression and classification. Regression is used when the target response Y we are trying to predict is continuous. Classification is used when the target response we are trying to predict belongs to a particular category or a finite class such as $Y \in \{1, 2, \ldots, k\}$. Figure 1 presents a generic workflow for supervised learning.

Figure 1: A generic workflow for supervised learning.

Example 1.1 (Image Classification)
Now suppose we are given a collection of images of either cats or dogs. We want to use this collection of images with labels to train a classifier (a prediction function) $\widehat{f}$ such that it can be used to classify a new image.

In addition to predictive analysis, another important goal of supervised learning is to do variable selection for better explanation.

1. Prediction: Given a new data point X, predict its response Y using $\widehat{Y} := \widehat{f}(X)$, where $\widehat{f}$ is constructed using training data such as in Example 1.1.

2. Better explanability : Find a small subset of $X_1, \ldots, X_d$ which is most related to Y.

2 Regression Analysis

Suppose we have $(Y, X) \sim P_{Y,X}$ at the population level, where $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$. We observe $(Y_1, X_1), \ldots, (Y_n, X_n) \sim P_{Y,X}$ at the sample level. Here $X_1, \ldots, X_n$ are i.i.d. copies of X. Denote $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ as the response vector and $\mathbf{X} = [X_{ij}] \in \mathbb{R}^{n \times d}$ as the design

matrix. We use (y_i, x_i) to denote the *realization* of (Y_i, X_i). Note that (Y_i, X_i) is random while (y_i, x_i) is deterministic.

Inference is the primary goal of statistics while prediction is the primary goal of maachine learning. We use the following example to highlight the differences between inference and prediction.

Example 2.1 (Inference and Prediction)
Suppose we have a population variable $X \sim \mathcal{N}(\theta, 1)$ with an unknown parameter $\theta \in \mathbb{R}$ and random samples $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$ with realizations $x_1, \ldots, x_n$.

1. For inference/statistics, we aim to infer the population quantities (e.g. θ). There are three main kinds of inference: point estimation $\widehat{\theta}_n(X_1, \ldots, X_n)$, confidence intervals of θ and hypothesis testing on θ (e.g. $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$).
2. For prediction/learning, we aim to predict a new X_{n+1} based on samples $X_1, \ldots, X_n$. In other words, generalization performance is the focus of learning.

Prior to the formulation of regression, we give an intuitive definition of regression.

Definition 2.1 (Regression)
Regression is the art of summarizing the relationship between two variables Y (response/label) and X (predictor/feature).

In other words, given the observed data $(Y_1, X_1), \ldots, (Y_n, X_n) \overset{\text{i.i.d.}}{\sim} P_{Y,X}$, we aim to find a prediction function f, such that $f(X)$ is "close" to Y. Recall that both X and Y are random. Naturally, we want to ask the following question.

Question 1. If $f(X)$ and Y are stochastic, how shall we characterize their "closeness"?

To answer this question, we need the definition of loss function and expected loss function. Let us first denote the space of possible data points by $\mathcal{X}$ and denote the space of possible labels by $\mathcal{Y}$.

Definition 2.2 (Loss Function)
A loss function $\ell(\cdot, \cdot)$ is a mapping $\ell : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}^+$. We will usually assume that ℓ is bounded below by some constant, typically 0.

Given the prediction function f and the loss function ℓ, the loss of an instance is $\ell(f(X), Y)$. We can then define the expected risk (or expected loss, or population risk)

$$R(f) := \mathbb{E}_{(Y,X)\sim P}[\ell(f(X), Y)].$$

Example 2.2 (Examples of Loss Functions)
We give some examples of (expected) loss functions.

1. ℓ_2-loss or ordinary least square (OLS) loss: $\ell(f(X), Y) = |f(X) - Y|^2$, $R(f) = \mathbb{E}_{(Y,X)\sim P}\{|f(X) - Y|^2\}$.
2. ℓ_1-loss or the least absolute deviation (LAD) loss: $\ell(f(X), Y) = |f(X) - Y|$, $R(f) = \mathbb{E}_{(X,Y)\sim P}\{|f(X) - Y|\}$.

The ℓ_2-loss is the most commonly used loss function because of its advantages:

1. ℓ_2 is mathematical simple;
2. ℓ_2 is computationally simple;
3. The minimizer of the expected loss $\mathbb{E}|f(X) - Y|^2$ is statistically justifiable as MLE under a Gaussian model, and MLE is the optimal estimator;
4. All smooth loss functions are locally quadratic by Taylor's expansion.

We have the following nice property for the ℓ_2-loss.

Theorem 2.3
Let $f^* = \operatorname{argmin}_f \mathbb{E}[f(X) - Y]^2$. Then we have

$$f^*(x) = \mathbb{E}[Y|X = x].$$

Proof. Let $\bar{f}(x) = \mathbb{E}[Y|X = x]$. Then we have the mean square error (MSE)

$$\begin{aligned}\mathbb{E}[Y - f(X)]^2 &= \mathbb{E}[Y - \bar{f}(X) + \bar{f}(X) - f(X)]^2 \\ &= \mathbb{E}[Y - \bar{f}(X)]^2 + \mathbb{E}[\bar{f}(X) - f(X)]^2 + 2\mathbb{E}[(Y - \bar{f}(X))(\bar{f}(X) - f(X))].\end{aligned}$$

For the third term, we have

$$\begin{aligned}\mathbb{E}[(Y - \bar{f}(X))(\bar{f}(X) - f(X))] &= \mathbb{E}_X\{\mathbb{E}[(Y - \bar{f}(X))(\bar{f}(X) - f(X))|X]\} \\ &= \mathbb{E}_X\{(\bar{f}(X) - f(X))\mathbb{E}[Y - \bar{f}(X)|X]\} \\ &= 0,\end{aligned}$$

where we used the double expectation rule $\mathbb{E}[Y] = \mathbb{E}_X\{\mathbb{E}[Y|X]\}$ in the first step and the fact that $\mathbb{E}[Y - \bar{f}(X)|X] = \mathbb{E}[Y|X] - \bar{f}(X) = 0$ in the last step.

Therefore, we have the decomposition $\mathbb{E}[Y - f(X)]^2 = \mathbb{E}[Y - \bar{f}(X)]^2 + \mathbb{E}[\bar{f}(X) - f(X)]^2$. The first term in the expansion is unavoidable, and thus it gives a lower bound on the loss we can achieve. The second term is nonnegative, so it reaches the minimum 0 when we take $f(x) = \bar{f}(x)$. That is, $\bar{f}(x)$ minimizes the MSE, i.e., $f^*(x) = \bar{f}(x)$. □

Remark 2.1
The $\mathbb{E}[Y|X = x] =: g(x)$ is the regression function/mean function that we aim to estimate/predict in the ℓ_2-regression analysis.

Considering that the expectation in $R(f)$ is taken with respect to the unknown distribution $P_{Y,X}$, we instead minimize the empirical risk

$$\widehat{R}(f) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2,$$

where $\widehat{R}(f)$ is the sample average approximation to $R(f)$. By minimizing $\widehat{R}(f)$, we have a good approximation of $f^* = \operatorname{argmin}_f R(f)$:

$$\widehat{f} := \operatorname*{argmin}_f \widehat{R}(f). \tag{2.1}$$

A trivial solution to (2.1) is an $\widehat{f}$ such that

$$\widehat{f}(x) = \begin{cases} Y_i, & \text{for } x = X_i \\ \text{anything}, & \text{for } x \neq X_1, \ldots, X_n. \end{cases}$$

However, such a construction would overfit the data since it can take anything when it is not evaluated at the training samples: the constructed function can be highly variable. We refer this phenomenon as the overfitting phenomenon.

Definition 2.4 (Overfitting)
A phenomenon when a statistical model has too much flexibility (or degrees of freedom) so that the model starts to predict the noise, rather than just predicting the signal.

Our solution to overfitting is regularization, by introducing additional information or constraint to reduce the flexibility or capacity of the model.

Example 2.3
Consider the OLS regression

$$\widehat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}[Y - f(X)]^2,$$

where $\mathcal{F}$ is referred to as a hypothesis (in the context of learning) or a model (in the context of statistics). By putting different constraints on f, we have many types of hypotheses $\mathcal{F}$. For examples,

$$\begin{aligned}
\text{Linear hypothesis: } & \mathcal{F} = \{f(x) : f(x) = \beta^\top x\}, \\
\text{Polynomial hypothesis: } & \mathcal{F} = \{f(x) : f(x) = \text{poly}(x)\}, \\
\text{Nonparametric hypothesis: } & \mathcal{F} = \{f(x) : \textstyle\int (f'(x))^2 \mathrm{d}x < \infty\}, \\
\text{Nonparametric hypothesis: } & \mathcal{F} = \{f(x) : f(x) \text{ is a residual network}\}.
\end{aligned}$$

Example 2.4 (Ordinary Least Square (OLS) Regression)
Given data $(Y_1, X_1), \ldots, (Y_n, X_n)$ with $X_i = (X_{i1}, \ldots, X_{id}) \in \mathbb{R}^d$ for $1 \le i \le n$. Denote $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ as the response vector and $\mathbf{X} = [X_{ij}] \in \mathbb{R}^{n \times d}$ as the design matrix with $(X_{11}, \ldots, X_{n1})^\top = (1, \ldots, 1)^\top$. Let $\widehat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^d} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ be the OLS estimator of β. Then $\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, because

$$\frac{\partial F(\beta)}{\partial \beta}\Big|_{\widehat{\beta}} = 0 \quad \Longleftrightarrow \quad 2\mathbf{X}^\top \mathbf{X}\widehat{\beta} - 2\mathbf{X}^\top \mathbf{Y} = 0,$$

where $F(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \mathbf{Y}^\top \mathbf{Y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta - 2\beta^\top \mathbf{X}^\top \mathbf{Y}$.

2.1 Model-based Interpretation of OLS

Consider the generative model

$$Y = \beta^\top X + \epsilon,$$

where $\mathbb{E}[\epsilon | X] = 0$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d}}{\sim} P_{X,Y}$. The joint log-likelihood is

$$\ell_n(\beta, \sigma^2) = \sum_{i=1}^n \log p_{\beta,\sigma^2}(Y_i, X_i) = \sum_{i=1}^n \log p_{\beta,\sigma^2}(Y_i | X_i) + \sum_{i=1}^n \log p(X_i),$$

where we used the fact that $p_{\beta,\sigma^2}(Y_i, X_i) = p_{\beta,\sigma^2}(Y_i | X_i) p(X_i)$ for $1 \le i \le n$. Then we have the MLE of β

$$\begin{aligned}
\widehat{\beta}^{\text{MLE}} &= \operatorname*{argmax}_\beta \ell_n(\beta, \sigma^2) \\
&= \operatorname*{argmin}_\beta \left\{ \sum_{i=1}^n \log p_{\beta,\sigma^2}(Y_i | X_i) + \sum_{i=1}^n \log p(X_i) \right\} \\
&= \operatorname*{argmin}_\beta \sum_{i=1}^n \log p_{\beta,\sigma^2}(Y_i | X_i) \\
&= \operatorname*{argmin}_\beta \left\{ \frac{1}{2} \sum_{i=1}^n \frac{(Y_i - \beta^\top X_i)^2}{\sigma^2} - n \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \right\} \\
&= \operatorname*{argmin}_\beta \sum_{i=1}^n \frac{(Y_i - \beta^\top X_i)^2}{2\sigma^2} \\
&= \operatorname*{argmin}_\beta \sum_{i=1}^n (Y_i - \beta^\top X_i)^2 \\
&= \operatorname*{argmin}_\beta \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \\
&= \text{The OLS estimator.}
\end{aligned}$$

Therefore, the solution to ordinary least squares regression is the MLE under a Gaussian noise model. Since the MLE is optimal in the sense that the MLE is the most efficient

estimator, the OLS estimator is also optimal under the Gaussian noise model. However, when the data is far away from a Gaussian noise model, ℓ_2-regression is not optimal more and probably should be avoided. Then we may ask the following question: Why do we care about model based interpretation? We list some reasons below.

1. We can know what kind of estimator is optimal under what models.
2. Based on the model/distribution information, we may construct confidence intervals, calculate p-values for certain hypotheses and evaluate the statistical significance.
3. The model-based approach provides a generative model, from which we can generate new samples.
4. We can provide Bayesian inference.

3 Classification

Classification is often referred to as discriminative analysis.

Definition 3.1 (Classification)
Classification is a supervised learning technique with a categorical response $Y \in \{C_1, \ldots, C_K\}$, where C_k, $1 \leq k \leq K$, denotes the label of the k-th class. Most of the time, we simply use $Y \in \{1, \ldots, K\}$. When $K = 2$, we call it **binary classification** and denote the two distinct classes by $\{-1, +1\}$, i.e., $Y \in \{-1, +1\}$. When $K > 2$, it is called **multiclass/multicategory classification** which can be reduced to binary classification problems if we consider a one-vs-one approach.

The goal of classification is to find a prediction mapping $h : X \to \{-1, +1\}$ such that the response Y and $h(X)$ are close enough to each other. Similar to regression analysis, we need to choose a loss function $\ell(\cdot, \cdot)$ and then take the expectation. Let us consider the commonly used ℓ_2 loss:

$$\ell(Y, h(X)) = |Y - h(X)|^2 = 4\mathbb{I}(Y \neq h(X)),$$

which becomes a 0-1 loss since $\mathbb{I}(Y \neq h(X))$ only takes two values 0 and 1. The 0-1 loss is a discrete loss often written as

$$\ell(Y, \widehat{Y}) = \begin{cases} 0, & Y = \widehat{Y}; \\ 1, & \text{otherwise.} \end{cases}$$

Recall our definition of expected loss (expected risk).

Definition 3.2 (Expected Loss / Expected Risk)
Given the prediction mapping $h(\cdot)$ and the loss function $\ell(\cdot, \cdot)$, we define the *expected loss* (or *expected risk*) as

$$R(h) := \mathbb{E}[\ell(Y, h(X))] = 4\mathbb{P}(Y \neq h(X)).$$

Definition 3.3 (Bayes Classification Rule)
The *Bayes classification rule* h^* is defined as

$$h^* = \operatorname*{argmin}_h R(h) = \operatorname*{argmin}_h \mathbb{P}(Y \neq h(X)).$$

We call $R^* := R(h^*)$ the *Bayes risk*.

The following theorem shows what the Bayes classification rule is.

Theorem 3.4 (Bayes Classification Rule)
The *Bayes classification rule* h^* is

$$h^*(x) = \begin{cases} +1, & \text{if } \mathbb{P}(Y = +1|X = x) > \frac{1}{2}; \\ -1, & \text{otherwise.} \end{cases}$$

Proof of Theorem 3.4.

$$\begin{aligned}
R(h) &= 4\mathbb{P}(Y \neq h(X)) \\
&\propto 1 - \mathbb{P}(Y = h(X)) \\
&= 1 - \sum_{y \in \{-1,+1\}} \mathbb{P}(Y = y, h(X) = y) \\
&= 1 - \sum_{y \in \{-1,+1\}} \mathbb{E}[\mathbb{I}(Y = y) \cdot \mathbb{I}(h(X) = y)] \\
&= 1 - \sum_{y \in \{-1,+1\}} \mathbb{E}_X \left[\mathbb{E}_{Y|X} \left\{(\mathbb{I}(Y = y) \cdot \mathbb{I}(h(X) = y)|X\right\}\right] \\
&= 1 - \sum_{y \in \{-1,+1\}} \mathbb{E}_X[\mathbb{I}(h(X) = y) \cdot \mathbb{P}(Y = y|X)] \\
&= 1 - \int_x [\mathbb{I}(h(x) = +1)\mathbb{P}(Y = +1|X = x) + \mathbb{I}(h(x) = -1)\mathbb{P}(y = -1|X = x)]\, p(x)\mathrm{d}x,
\end{aligned}$$

where $p(x)$ is the density function of X. We want to minimize the risk by maximizing the integral in the last line. Consider the two terms in the integrand, $\mathbb{I}(h(x) = +1)\mathbb{P}(Y = +1|X = x)$ and $\mathbb{I}(h(x) = -1)\mathbb{P}(y = -1|X = x)$. For any x, there must be one term being positive and the other is 0. So we determine the value of $h^*(x)$ such that

$$\begin{aligned}
&\mathbb{I}(h^*(x) = +1)\mathbb{P}(Y = +1|X = x) + \mathbb{I}(h^*(x) = -1)\mathbb{P}(y = -1|X = x) \\
&= \max\{\mathbb{P}(Y = +1|X = x), \mathbb{P}(y = -1|X = x)\}.
\end{aligned}$$

That is,

$$h^*(x) = \begin{cases} +1, & \text{if } \mathbb{P}(Y = +1|X = x) > \frac{1}{2}; \\ -1, & \text{otherwise.} \end{cases}$$

□

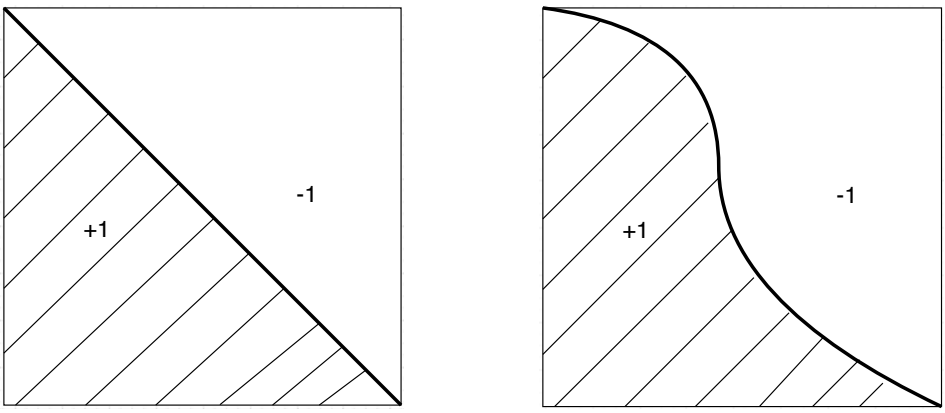

Figure 2: A linear decision boundary and a nonlinear decision boundary respectively.

3.1 Decision Boundary

By the theorem on Bayes rule, the key of classification is to model $\mathbb{P}(Y = +1|X = x) =: r(x)$ so that we can estimate/approximate it.

Definition 3.5 (Decision Boundary)
We define *decision boundary* as the function

$$D(r) = \left\{ x : r(x) = \frac{1}{2} \right\}.$$

The decision boundary, corresponding to a classifier, is the boundary of two regions: positive regions and negative regions. If an instance falls in the positive region, the classifier will classify this instance to the positive class; vice versa. If $r(x)$ is a linear (nonlinear) function of x then we say it has a linear (nonlinear) decision boundary. The corresponding classifier is then called a linear (nonlinear) classifier. Figure 2 presents an illustration of the linear and nonlinear decision boundary.

The conditional probability $\mathbb{P}(Y = +1|X = x)$ can be modeled in various ways. The most commonly used modeling is the logistic modeling.

Example 3.1 (Logistic Modeling)
We use the logistic function to model the conditional probability

$$\mathbb{P}(Y = +1|X = x) = \frac{1}{1 + e^{-f(x)}},$$

or equivalently,

$$\mathbb{P}(Y = -1|X = x) = \frac{1}{1 + e^{f(x)}},$$

where f is the parameter of interest. We can also write this model in a more compact way.

$$\mathbb{P}(Y = y|X = x) = \frac{1}{1 + e^{-yf(x)}},$$

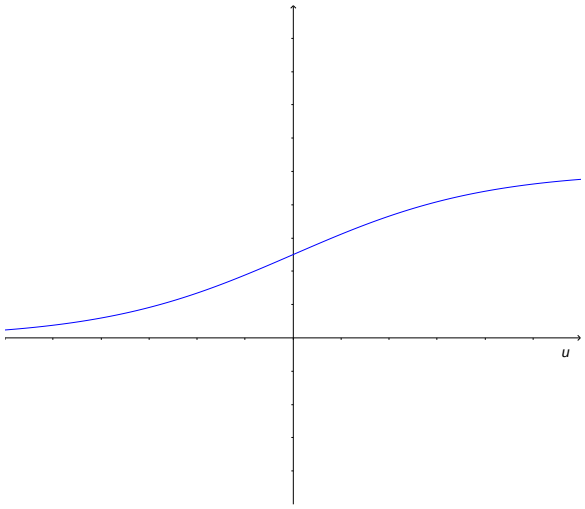

Figure 3: Logistic function: $g(u) = \frac{1}{1+e^{-u}}$.

where $yf(x)$ is referred to as the *margin function* (or simply the *margin*). Figure 3 presents the geometry of the logistic function.

We can also use the probit modeling.

Example 3.2 (Probit Modeling)
Probit modeling uses the CDF of the standard normal distribution to model the conditional probability

$$\mathbb{P}(Y = +1|X = x) = \Phi(f(x)),$$

where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0, 1)$ and f is the parameter of interest.

Remark 3.1
The logistic modeling is more "principled" and prefered.

3.2 The Statistical Model of Logistic Regression and Its MLE

We consider the following statistical model for logistic regression

$$\mathcal{P} := \{p(y, x) = p_f(Y = y|X = x) \cdot p_X(x) : \text{ for all possible } f, p_X(x)\},$$

where f is our parameter of interest. Note that p_X is a nuisance parameter since we only care the conditional probability $\mathbb{P}(Y = y|X = x)$ for classification problem. Given samples

(Y_i, X_i), $1 \le i \le n$, we have the likelihood function

$$L_n(f) = \prod_{i=1}^{n} p_f(Y_i|X_i) \cdot p_X(X_i)$$

and the log-likelihood function

$$\begin{aligned}\ell_n(f) &= \prod_{i=1}^{n} p_f(Y_i|X_i) p_X(X_i) \\ &= \sum_{i=1}^{n} \log p_f(Y_i|X_i) + \underbrace{\sum_{i=1}^{n} \log p_X(X_i)}_{\text{does not dependent on } f}.\end{aligned}$$

Then the MLE is

$$\begin{aligned}\widehat{f} &= \operatorname*{argmax}_f \ell_n(f) \\ &= \operatorname*{argmax}_f \sum_{i=1}^{n} \log p_f(Y_i|X_i) \\ &= \operatorname*{argmin}_f \sum_{i=1}^{n} \log(1 + e^{-Y_i f(X_i)}).\end{aligned}$$

Remark 3.2 (Overfitting)
If we do not impose any constraint on f, then a trivial solution is

$$\widehat{f}(x) = \begin{cases} +\infty, & \text{if } X_i = x, Y_i = +1; \\ -\infty, & \text{if } X_i = x, Y_i = -1; \\ \text{arbitrary}, & \text{elsewhere.} \end{cases}$$

A solution to prevent overfitting is to reduce model complexity by adding constraints. Some typical constraints are shown in the following examples.

Example 3.3 (Linear Logistic Regression)
We restrict f to the linear function class

$$f \in \mathcal{F} := \left\{ f(x) : f(x) = \beta_0 + \beta^\top x \right\}.$$

Example 3.4 (Nonparametric Logistic Regression)
We assume that f is continuous and has square integrable second derivative, i.e.,

$$f \in \mathcal{F} := \left\{ f(x) : f(x) \text{ is continuous and } \int [f''(x)]^2 < \infty \right\}.$$

3.3 A Risk Minimization Interpretation of Logistic Regression

Besides the model-based approach that naturally leads to the logistic regression, we now give a risk-based interpretation. This also allows for various generalizations. We first define a loss function that corresponds the logistic regression.

Definition 3.6 (Logistic Loss)
The logistic regression induces a loss function

$$\ell^{\text{logistic}}(y, f(x)) := \log(1 + e^{-yf(x)}),$$

where we encourage the term $yf(x)$ to be larger. Figure 4 presents the shape of the logistic loss.

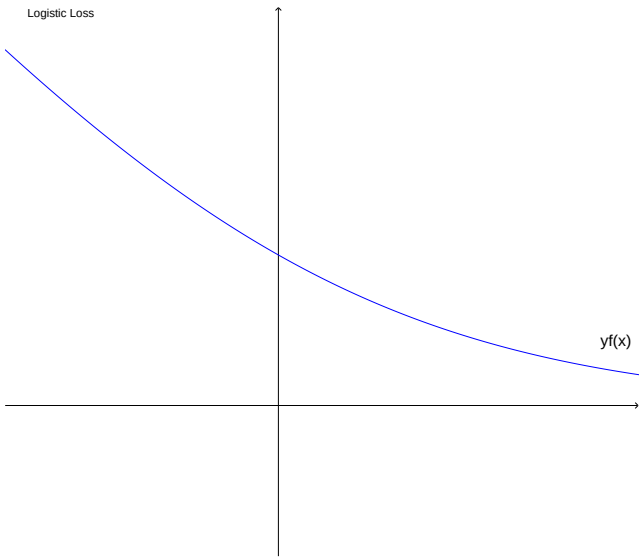

Figure 4: Logistic loss: $\log(1 + e^{-u})$.

Definition 3.7 (Logistic Risk)
The expected loss based the logistic loss function is

$$R(f) = \mathbb{E}\left[\log\left(1 + e^{-yf(x)}\right)\right].$$

To minimize the logistic risk $R(f)$, we encourage y and $f(x)$ to have the same sign. Recall the logistic modeling in Example 3.1:

$$\mathbb{P}(Y = +1|X = x) = \frac{1}{1 + e^{-f(x)}} = \frac{e^{f(x)}}{1 + e^{f(x)}},$$
$$\mathbb{P}(Y = -1|X = x) = \frac{1}{1 + e^{f(x)}},$$

we have the log odds ratio

$$f(x) = \log\left(\frac{\mathbb{P}(Y = +1|X = x)}{\mathbb{P}(Y = -1|X = x)}\right).$$

When $f(x) > 0$, it is more likely to have that $Y = +1$.

Remark 3.3
When $Y = +1$, we hope the log-odds ratio to be big. When $Y = -1$, we hope it to be small.

3.4 Other Loss Functions

By viewing logistic regression from the ERM perspective, we can consider different loss functions which lead to various methods for classification.

Some more examples using other loss functions are shown below.

Example 3.5 (Quadratic loss)
We consider the ℓ_2 loss, which leads to the ℓ_2 regression

$$\widehat{\beta}^{\lambda} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta^\top X_i)^2 = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (1 - Y_i \beta^\top X_i)^2$$

Example 3.6 (0-1 loss)
The 0-1 loss function is defined as

$$\mathbb{I}(\text{Y has a different sign as } \beta^\top X) = \mathbb{I}(Y f(X) < 0).$$

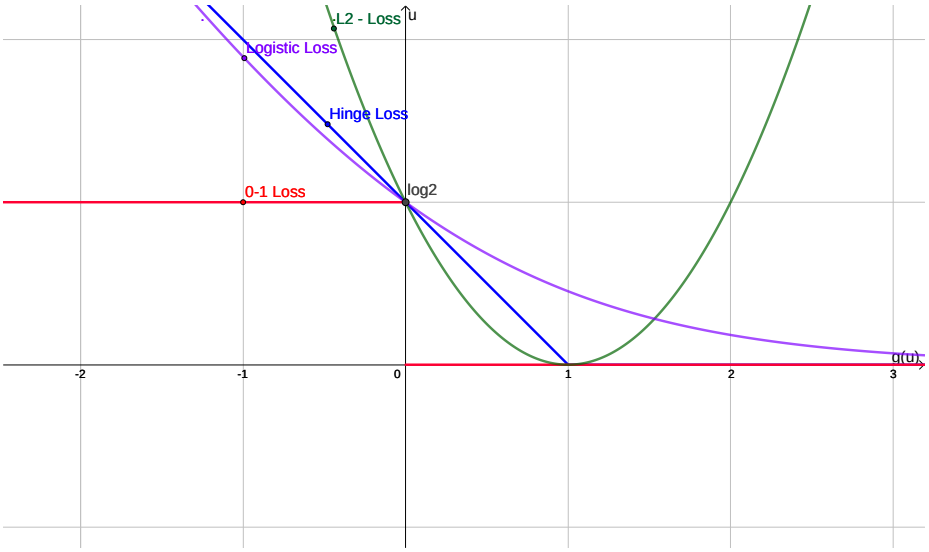

Figure 5: Four types of loss functions: logistic loss, ℓ_2 loss, 0-1 loss and hinge loss.

Example 3.7 (Hinge loss and support vector machine)
We use the *hinge loss* $g(u) = (1-u)_+$ for fitting and ℓ_2-norm for penalty

$$\widehat{\beta}^{\lambda} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (1 - Y_i \beta^\top X_i)_+ + \lambda \|\beta\|_2^2 \right\},$$

where the penalty $\lambda\|\beta\|_2^2$ is added to penalize the model complexity.

Remark 3.4
From a statistical perspective, the logistic modeling is very natural as it corresponds to the maximum likelihood estimator. From the risk minimization perspective, we can consider various losses such as ℓ_2-loss, 0/1-loss, logistic loss, and hinge loss,