Lecture-11: Multi-class classification

1 Introduction

We will consider the following two classes of algorithms.

1. *Uncombined algorithms*: Specifically designed for the multi- class setting such as multi-class SVMs, decision trees, or multi-class boosting.
2. *Aggregated algorithms*: Based on reduction to binary classification and require training multiple binary classifiers.

As before, we will denote the input space by $\mathcal{X}$ the output space by $\mathcal{Y}$, and unknown distribution by $\mathcal{D} \in \mathcal{M}(\mathcal{X})$ over input space $\mathcal{X}$ according to which input points are drawn. We will consider the following two multi-class cases.

1. *Mono-label case:* The output space $\mathcal{Y}$ is a finite set of classes marked $\mathcal{Y} = \{0, ..., M-1\}$ without any loss of generality. Each example in this case, is labeled with a single class.
2. *Multi-label case:* The output space $\mathcal{Y} = \{-1, +1\}^k$ is binary vector. Each example in this case, can be labeled with several classes. The positive components of a vector in $\{-1, +1\}^k$ indicate the classes associated with an example. For example, text documents can be labeled with several different relevant topics, e.g., sports, business, and society.

The learner receives a labeled sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ with $x \in \mathcal{X}^m$ drawn *i.i.d.* according to $\mathcal{D}$, and $y_i = c(x_i)$ for all $i \in [m]$, where $c : \mathcal{X} \to \mathcal{Y}$ is the true concept. Thus, we consider a deterministic scenario, which can be straightforwardly extended to a stochastic one that admits a distribution over $\mathcal{X} \times \mathcal{Y}$.

Definition 1.1 (zero-one loss function). We define **zero-one loss function** $d : \mathcal{Y} \times \mathcal{Y}$ for a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ as

$$d(h(x), y) \triangleq \mathbb{1}_{\{h(x) \neq y\}}.$$

Definition 1.2 (Hamming distance). We define **Hamming distance** $d_H : \mathcal{Y} \times \mathcal{Y}$ for a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ for any output space $\mathcal{Y} \subseteq R^k$ as

$$d_H(h(x), y) \triangleq \sum_{\ell=1}^{k} \mathbb{1}_{\{h(x)_\ell \neq y_\ell\}}.$$

Remark 1. Empirical error for any loss function d, hypothesis h, and labeled sample S, is given as $\hat{R}(h) = \frac{1}{m}\sum_{i=1}^{m} d(h(x_i), y_i)$.

Challenges in multi-class setting.

1. Computational challenges for large M, k
2. Unbalanced classes, and poor performance guarantees on classes with small training sample, and large generalization error due to classes with large training sample
3. Hierarchical relationship between classes

2 Bayesian framework

We assume that the *i.i.d.* sample comes from a known distribution $\mathcal{D}_y$ if the data has label $y \in \mathcal{Y}$. We further assume that a prior probability distribution on data coming from each class is $\pi \in \mathcal{M}(\mathcal{Y})$. For a hypothesis $h : \mathcal{X} \to \mathcal{Y}$, the loss for a labeled example (x_i, y_i) is given by $d(h(x_i), y_i)$. We observe that each hypothesis $h : \mathcal{X} \to \mathcal{Y}$ is equivalently characterized by the partition of input spaces $\mathcal{X}$ given by $(E_y, y \in \mathcal{Y})$ where $E_y \triangleq \{x \in \mathcal{X} : h(x) = y\} = h^{-1}\{y\}$.

Definition 2.1 (Bayesian loss). Bayesian loss function $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ for a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ and a labeled samples (x, y) is defined as

$$d(h(x), y) \triangleq \sum_{z \in \mathcal{Y}} c_{zy} \mathbb{1}_{\{h(x)=z\}} = \sum_{z \in \mathcal{Y}} c_{zy} \mathbb{1}_{E_z}(x).$$

We assume the cost of correct decision is smaller than incorrect decisions and hence $c_{yy} < c_{zy}$ for all possible classes $y, z \in \mathcal{Y}$.

Definition 2.2 (Bayes risk). Bayes risk $R : \mathcal{Y}^{\mathcal{X}} \to \mathbb{R}_+$ is defined for each hypothesis $h \in \mathcal{X} \to \mathcal{Y}$ as

$$R(h) \triangleq \mathbb{E}[d(h(X), c(X))],$$

for Bayesian loss function d and sample X with prior probability distribution of π on being from one of the M classes, and distribution $\mathcal{D}_y$ for a sample with label $y \in \mathcal{Y}$.

Problem 1. Find the Bayesian optimal hypothesis h that minimizes the Bayesian risk R.

Remark 2. Denoting the density of example x with label y as $\frac{d\mathcal{D}_y}{dx} = f(x \mid H_y)$, we can write the density of example x as $f(x) \triangleq \frac{\sum_{y \in \mathcal{Y}} d\mathcal{D}_y(x)\pi_y}{dx}$. Defining the conditional probability of label y given example x as $P(H_y \mid x) \triangleq \frac{d\mathcal{D}_y(x)\pi_y}{f(x)dx}$, we can write the infinitesimal probability of example x being generated from class y as

$$d\mathcal{D}_y(x)\pi_y = f(x \mid H_y)\pi_y dx = f(x)P(H_y \mid x)dx.$$

Defining the mean cost of hypothesis h declaring label z for an example x as $c_z(x) \triangleq \sum_{y \in \mathcal{Y}} c_{zy} P(H_y \mid x)$, we can write the Bayes risk as

$$R(h) = \sum_{y \in \mathcal{Y}} \pi_y \sum_{z \in \mathcal{Y}} c_{zy} \int_{x \in \mathcal{X}} \mathbb{1}_{E_z}(x) f(x \mid H_y) dx = \int_{x \in \mathcal{X}} dx f(x) \sum_{z \in \mathcal{Y}} \mathbb{1}_{E_z}(x) \sum_{y \in \mathcal{Y}} c_{zy} P(H_y \mid x) = \int_{x \in \mathcal{X}} dx f(x) \sum_{z \in \mathcal{Y}} \mathbb{1}_{E_z}(x) c_z(x).$$

Finding the Bayes optimal hypothesis is identical to finding regions $(E_z, z \in \mathcal{Y})$ such that the cost $c_z(x)$ is minimum for each $x \in \mathcal{X}$. That is, we find

$$E_z \triangleq \left\{ x \in \mathcal{X} : c_z(x) = \min_{w \in \mathcal{Y}} c_w(x) \right\}.$$

Definition 2.3. The Bayes optimal hypothesis is $h_B(x) \triangleq \arg\min \{c_w(x) : w \in \mathcal{Y}\}$.

2.1 Special Bayesian loss case

We consider the special Bayesian loss case when cost of correct classification is zero, and incorrect classification is unity for all incorrect classifications. That is, $d(h(x), y) = \mathbb{1}_{\{y \neq h(x)\}}$ for any hypothesis h and hence $c_{zy} = \mathbb{1}_{\{z \neq y\}}$ for all labels $z, y \in \mathcal{Y}$. For this case, the mean cost of hypothesis h declaring label z for an example x is

$$c_z(x) = \sum_{y \in \mathcal{Y}} c_{zy} P(H_y \mid x) = \sum_{y \neq z} P(H_y \mid x) = 1 - P(H_z \mid x).$$

Remark 3. For the zero-one loss, the optimal Bayesian hypothesis is $h_B(x) \triangleq \arg\max \{P(H_z \mid x) : z \in \mathcal{Y}\}$ the one that maximizes the *a posteriori* probability of observing a label given an example x.

Definition 2.4. The hypothesis that maximizes the *a posteriori* probability of observing a label given an example x, is called a **MAP hypothesis** and is given by

$$h_{\text{MAP}}(x) \triangleq \arg\max \{P(H_z \mid x) : z \in \mathcal{Y}\} = \arg\max \{f(x \mid H_z)\pi_z : z \in \mathcal{Y}\}.$$

Remark 4. If all labels are equally likely *a priori*, then $h_{\text{MAP}}(x) = \arg\max \{f(x \mid H_z) : z \in \mathcal{Y}\}$ that is the hypothesis maximizes the likelihood of observing example x.

Definition 2.5. The hypothesis that maximizes the likelihood of observing an example x, is called an **ML hypothesis** and is given by

$$h_{\text{ML}}(x) \triangleq \arg\max \{f(x \mid H_z) : z \in \mathcal{Y}\}.$$

Example 2.6 (Gaussian distribution). Consider the multi-class classification case when the output space $\mathcal{Y} = \{0, \dots, M-1\}$, and the density $f(x \mid H_y)$ of example $x \in \mathcal{X} = \mathbb{R}^d$ for label $y \in \mathcal{Y}$ is a Gaussian distribution with mean vector $\mu_y \in \mathbb{R}^d$ and variance matrix K_y. The maximum likelihood (ML) classifier is given by

$$h_{\text{ML}}(x) = \arg\max \left\{ \exp\Big(-\frac{1}{2}(x-\mu_y)^T K_y^{-1}(x-\mu_y) \Big) : y \in \mathcal{Y} \right\}.$$

When $K_y = \sigma^2 I$ for all $y \in \mathcal{Y}$, we get

$$h_{\text{ML}}(x) = \arg\max \left\{ -\left\| x - \mu_y \right\|^2 : y \in \mathcal{Y} \right\} = \arg\min \left\{ \left\| x - \mu_y \right\| : y \in \mathcal{Y} \right\}.$$

This is called the **minimum distance classifier**.

Example 2.7 (Communication over Gaussian channels). Consider a communication channel with additive white Gaussian noise pair $N : \Omega \to \mathbb{R}^2$ with independent components having mean zero and variance σ^2. For an input pair $y \in (\{0,1\}^2$, the output pair $x \in \mathbb{R}^2$ is given by $(x^1, x^2) = (y^1 + N^1, y^2 + N^2)$. Given the output x, one wants to classify input y. The minimum distance classifier gives for each output $x \in \mathbb{R}^2$,

$$h(x) \triangleq \arg\min \left\{ (x^1 - y^1)^2 + (x^2 - y^2)^2 : (y^1, y^2) \in \{0,1\} \times \{0,1\} \right\}.$$

3 Machine learning framework

We may know the distribution $\mathcal{D}_y$ for each label $y \in \mathcal{Y}$. Though, we may know or assume the prior distribution π. The *posterior* distribution given a labeled sample S of m *i.i.d.* examples, is defined as

$$Q_y \triangleq P(H_y \mid S).$$

For this measure, we can write the loss function as

$$R(h) \triangleq \mathbb{E}_Q d(h(x), y).$$

Theorem 3.1. *With probability greater than* $1 - \delta$, *we have*

$$R(h) \leqslant \hat{R}(h) + \Big(\frac{D(Q \| P) + \ln \frac{m}{\delta}}{2m - 1} \Big)^{\frac{1}{2}},$$

where KL distance $D(Q\|P) \triangleq \sum_{x \in \mathcal{X}} Q(x) \ln \frac{Q(x)}{P(x)}$ *if* $\text{supp}(Q) \subseteq \text{supp}(P)$ *and infinite otherwise.*