EE6550 Machine Learning

Lecture Seven – Multi-Class Classification

Chung-Chin Lu

Department of Electrical Engineering

National Tsing Hua University

April 24, 2017

Motivations

- Real-world problems often have multiple classes: documents, speeches, images, biological sequences.
- Algorithms studied so far: designed for binary classification problems.
- How do we design multi-class classification algorithms?
  - Can the algorithms used for binary classification be extended to multi-class classification?
  - Can we reduce a multi-class classification problem to multiple binary classification problems?

The Contents of This Lecture

- Multi-class classification problem.
- Generalization bound.
- Uncombined multi-class algorithms.
- Aggregated multi-class algorithms.

Multi-Class Classification Problem

- Training data : a sample $S = (\omega_1, \ldots, \omega_m)$ of size m drawn i.i.d. from an input space $\mathscr{I}$ according to some fixed but unknown distribution D with labels $(c(\omega_1), \ldots, c(\omega_m))$ from a fixed but unknown concept c.
  - Mono-label case : the label space is $\mathscr{Y} = \{1, 2, \ldots, k\}$.
  - Multi-label case : the label space is $\mathscr{Y} = \{-1, +1\}^k$.
- Problem : finding a classifier $h_S : \mathscr{I} \to \mathscr{Y}$ in the hypothesis set $\mathcal{H}$ with small generalization error by training the learning algorithm with the labeled sample S.
  - Mono-label case : $R(h_S) = \underset{\omega \sim D}{E}[1_{h_S(\omega) \neq c(\omega)}]$.
  - Multi-label case : $R(h_S) = \underset{\omega \sim D}{E}[\frac{1}{k} \sum_{i=1}^{k} 1_{h_S(\omega)_i \neq c(\omega)_i}] = \underset{\omega \sim D}{E}[\frac{1}{k} d_H(h_S(\omega), c(\omega))]$, where d_H is the Hamming distance.

Remarks

- In most tasks considered, the number k of classes is ≤ 100.
- For large k, the problem is often not treated as a multi-class classification problem (ranking or density estimation, e.g., automatic speech recognition).
  - Computational efficiency issues arise for larger k's.
- In general, classes are not balanced.
  - Some classes may be represented by less than 5 percent of the labeled sample, while others may dominate a very large fraction of the data.

- When separate binary classifiers are used to define the multi-class solution, we may need to train a classifier distinguishing between two classes with only a small representation in the training sample. This implies training on a small sample, with poor performance guarantees.
- Alternatively, when a large fraction of the training instances belong to one class, it may be tempting to propose a hypothesis always returning that class, since its generalization error as defined earlier is likely to be relatively low. However, this trivial solution is typically not the one intended.
- Instead, the loss function may need to be reformulated by assigning different misclassification weights to each pair of classes.

- The relationship between classes may be hierarchical.
  - For example, in the case of document classification, the error of misclassifying a document dealing with world politics as one dealing with real estate should naturally be penalized more than the error of labeling a document with sports instead of the more specific label baseball.
  - Thus, a more complex and more useful multi-class classification formulation would take into consideration the hierarchical relationships between classes and define the loss function in accordance with this hierarchy.
  - More generally, there may be a graph relationship between classes as in the case of the GO ontology in computational biology.
  - The use of hierarchical relationships between classes leads to a richer and more complex multi-class classification problem.

The Contents of This Lecture

- Multi-class classification problem.
- Generalization bound.
- Uncombined multi-class algorithms.
- Aggregated multi-class algorithms.

Multi-Class Classifiers – Mono-Label Case

- In the binary setting, a classifier (a hypothesis) $h : \mathscr{I} \to \mathscr{Y}$ is often defined based on the sign of a scoring function $\tilde{h} : \mathscr{I} \to \mathbb{R}$, i.e.,

$$h(\omega) = \operatorname{sgn}(\tilde{h}(\omega)) \;\forall\; \omega \in \mathscr{I}.$$

- In the multi-class setting, a classifier (a hypothesis) $h : \mathscr{I} \to \mathscr{Y}$ is defined based on a scoring function $\tilde{h} : \mathscr{I} \times \mathscr{Y} \to \mathbb{R}$ such that the label of an item ω in the input space $\mathscr{I}$ predicted by h is

$$h(\omega) \triangleq \arg\max_{y \in \mathscr{Y}} \tilde{h}(\omega, y).$$

  - There is an arbitration if there are more than one $y \in \mathscr{Y}$ which reaches the maximum value of $\tilde{h}(\omega, \cdot)$.

Margin of a Multi-Class Classifier – Mono-Label Case

- Margin : the margin of the scoring function $\tilde{h}$ at a labeled item $(\omega, c(\omega))$ is defined as

$$\rho_{\tilde{h}}(\omega, c(\omega)) = \tilde{h}(\omega, c(\omega)) - \max_{y \in \mathscr{Y}, y \neq c(\omega)} \tilde{h}(\omega, y).$$

  - The classifier h misclassifies item ω only if $\rho_{\tilde{h}}(\omega, c(\omega)) \leq 0$.

- Empirical ρ-margin loss : for each $\rho > 0$, the empirical ρ-margin loss of a hypothesis h for multi-class classification w.r.t. the concept c on the labeled sample $S = (\omega_1, \ldots, \omega_m)$ of size m is defined as

$$\hat{R}_{S,\rho}(h) \triangleq \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(\rho_{\tilde{h}}(\omega_i, c(\omega_i))),$$

where

$$\Phi_\rho(x) = \begin{cases} 1, & \text{if } x \le 0, \\ 1 - \frac{x}{\rho}, & \text{if } 0 \le x \le \rho, \\ 0, & \text{if } x \ge \rho \end{cases}$$

is the ρ-margin loss function.

- $\hat{R}_{S,\rho}(h)$ is upper bounded by the fraction of the training items misclassified by h or correctly classified but with margin less than or equal to ρ:

$$\hat{R}_S(h) \le \frac{1}{m} \sum_{i=1}^{m} 1_{\rho_{\tilde{h}}(\omega_i, c(\omega_i)) \le 0} \le \hat{R}_{S,\rho}(h) \le \frac{1}{m} \sum_{i=1}^{m} 1_{\rho_{\tilde{h}}(\omega_i, c(\omega_i)) \le \rho}.$$

A Lemma

Lemma 8.1: Let

- $\mathcal{H}_1, \ldots, \mathcal{H}_l$: l hypothesis sets, each consisting of measurable functions from the input space $\mathscr{I}$ to the output space $\mathscr{Y}' \subseteq \mathbb{R}$;
  - Assume that $\sup_{h_i \in \mathcal{H}_i} |h_i(\omega)| < +\infty$ for all $\omega \in \mathscr{I}$ and for all $i \in [1, l]$.
- $\mathcal{G} = \{\max(h_1, \ldots, h_l) \mid h_i \in \mathcal{H}_i,\ i \in [1, l]\}$.

Then, for any sample S of size m, the empirical Rademacher complexity of $\mathcal{G}$ can be upper bounded as follows:

$$\hat{\mathfrak{R}}_S(\mathcal{G}) \leq \sum_{j=1}^{l} \hat{\mathfrak{R}}_S(\mathcal{H}_j).$$

Proof.

- $S = (\omega_1, \ldots, \omega_m)$: a sample of size m.
- When $l = 2$, $\max(h_1, h_2) = \frac{1}{2}(h_1 + h_2 + |h_1 - h_2|)$ for all $h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2$.

Thus we have

$$\begin{aligned}
& \hat{\mathfrak{R}}_S(\mathcal{G}) \\
= \quad & \frac{1}{2^m} \sum_{\sigma_1, \sigma_2, \ldots, \sigma_m \in \{-1,+1\}} \sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(\omega_i) \\
= \quad & \mathop{E}_{\sigma} \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(\omega_i) \right] \text{ where } \sigma = (\sigma_1, \ldots, \sigma_m) \text{ is a random} \\
& \text{vector with uniform distribution over } \{-1, +1\}^m \\
= \quad & \mathop{E}_{\sigma} \left[\sup_{h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \max(h_1(\omega_i), h_2(\omega_i)) \right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2} \mathop{E}_{\sigma}\left[\sup_{h_1\in\mathcal{H}_1, h_2\in\mathcal{H}_2} \frac{1}{m}\sum_{i=1}^{m} \sigma_i (h_1(\omega_i) + h_2(\omega_i) + |(h_1 - h_2)(\omega_i)|)\right] \\
&\leq \frac{1}{2} \mathop{E}_{\sigma}\left[\sup_{h_1\in\mathcal{H}_1} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h_1(\omega_i) + \sup_{h_2\in\mathcal{H}_2} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h_2(\omega_i)\right. \\
&\qquad\qquad \left. + \sup_{h_1\in\mathcal{H}_1, h_2\in\mathcal{H}_2} \frac{1}{m}\sum_{i=1}^{m} \sigma_i |(h_1 - h_2)(\omega_i)|\right]
\end{aligned}
$$

by the subadditivity of sup, i.e., $\sup_i(u_i + v_i) \leq \sup_i u_i + \sup_i v_i$. Since $||u| - |v|| \leq |u - v| \ \forall\ u, v \in \mathbb{R}$ implies that the mapping $u \mapsto |u|$ is a 1-Lipschitz function from $\mathbb{R}$ to $\mathbb{R}$. Since $\sup_{h_i\in\mathcal{H}_i} |h_i(\omega)| < +\infty$ for all $\omega \in \mathscr{I}$ and for all $i = 1, 2$, we have

$$
\sup_{h_1\in\mathcal{H}_1, h_2\in\mathcal{H}_2} \left(\sum_{i=1}^{j} \sigma_i |(h_1 - h_2)(\omega_i)| + \sum_{i=j+1}^{m} \sigma_i (h_1 - h_2)(\omega_i)\right) < +\infty
$$

for all $\sigma_i \in \{-1, +1\}, i \in [1, m]$, for all $j \in [0, m]$ and for all samples

$S = (\omega_1, \ldots, \omega_m)$ of size m. Now we have

$$\begin{aligned}
& \mathop{E}_{\sigma}\left[\sup_{h_1\in\mathcal{H}_1, h_2\in\mathcal{H}_2} \frac{1}{m}\sum_{i=1}^{m} \sigma_i |(h_1 - h_2)(\omega_i)|\right] \\
\leq\ & \mathop{E}_{\sigma}\left[\sup_{h_1\in\mathcal{H}_1, h_2\in\mathcal{H}_2} \frac{1}{m}\sum_{i=1}^{m} \sigma_i (h_1 - h_2)(\omega_i)\right]
\end{aligned}$$

by applying Talagrand's lemma in Lecture 3 to the hypothesis set $\mathcal{H} = \mathcal{H}_1 - \mathcal{H}_2$ and the 1-Lipschitz function $\Phi = |\cdot|$

$$\leq \mathop{E}_{\sigma}\left[\sup_{h_1\in\mathcal{H}_1} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h_1(\omega_i) + \sup_{h_2\in\mathcal{H}_2} \frac{1}{m}\sum_{i=1}^{m} (-\sigma_i) h_2(\omega_i)\right]$$

again by the subadditivity of sup. We conclude that

$$\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{G}) &\leq \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H}_1) + \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H}_2) + \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H}_1) + \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H}_2) \\
&= \hat{\mathfrak{R}}_S(\mathcal{H}_1) + \hat{\mathfrak{R}}_S(\mathcal{H}_2).
\end{aligned}$$

- For the general case of $l \geq 2$, we repeatedly use the case of $l = 2$ by noting that

$$\max(h_1, \ldots, h_l) = \max(h_1, \max(h_2, \ldots, h_l))$$

and

$$\sup_{h_2 \in \mathcal{H}_2, \ldots, h_l \in \mathcal{H}_l} |\max(h_2, \ldots, h_l)(\omega)| = \max_{j \in [2,l]} \sup_{h_j \in \mathcal{H}_j} |h_j(\omega)| < +\infty$$

for all $\omega \in \mathscr{I}$.

This proves the lemma. $\square$

Margin Bound for Multi-Class Classification – Mono-Label Case

Theorem 8.1: Let

- $\mathscr{I}$: the input space of all possible items ω, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$.
- $c : \mathscr{I} \to \mathscr{Y} = \{1, 2, \ldots, k\}$: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $\mathcal{H}$: a set of hypotheses h defined based on corresponding measurable scoring functions $\tilde{h} : \mathscr{I} \times \mathscr{Y} \to \mathbb{R}$.
  - Assume that $\sup_{h \in \mathcal{H}} |\tilde{h}(\omega, y)| < +\infty$ for all $\omega \in \mathscr{I}$ and for all $y \in \mathscr{Y}$.
- $\Pi_1(\mathcal{H}) = \{\omega \mapsto \tilde{h}(\omega, y) \mid h \in \mathcal{H}, y \in \mathscr{Y}\}$.
  - Members of $\Pi_1(\mathcal{H})$ are measurable functions from $\mathscr{I}$ to $\mathbb{R}$.

- $S = (\omega_1, \ldots, \omega_m)$: a labeled sample of size m with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- $\rho > 0$: a given margin.
- $\hat{R}_{S,\rho}(h) \triangleq \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(\rho_{\tilde{h}}(\omega_i, c(\omega_i)))$: the empirical ρ-margin loss of the hypothesis h for multi-class classification w.r.t. the concept c on a labeled sample S of size m.

Then for any $\delta > 0$, with probability at least $1 - \delta$, the following multi-class classification generalization bound holds for all h in $\mathcal{H}$:

$$
\begin{aligned}
R(h) &\leq \hat{R}_{S,\rho}(h) + \frac{2k^2}{\rho}\mathfrak{R}_m(\Pi_1(\mathcal{H})) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}, \\
R(h) &\leq \hat{R}_{S,\rho}(h) + \frac{2k^2}{\rho}\hat{\mathfrak{R}}_S(\Pi_1(\mathcal{H})) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.
\end{aligned}
$$

Proof.

- $g_h : \mathscr{I} \times \{1, 2, \ldots, k\} \to [0, 1]$: the ρ-margin loss function associated with the hypothesis h, defined as

$$g_h(\omega, y) \triangleq \Phi_\rho(\rho_{\tilde{h}}(\omega, y)).$$

- $\mathcal{G} = \{g_h \mid h \in \mathcal{H}\}$: the family of ρ-margin loss functions associated with hypotheses in $\mathcal{H}$.
- $\mathscr{Z} = \mathscr{I} \times \{1, 2, \ldots, k\}$: the input set of ρ-margin loss functions g_h, associated with a probability space $(\mathscr{Z}, \tilde{\mathcal{F}}, \tilde{P})$ where $\tilde{P}$ is an extension of P from on $\mathcal{F}$ to on $\tilde{\mathcal{F}} = \mathcal{F} \times 2^{\{1,2,\ldots,k\}}$.
- $\tilde{S} = ((\omega_1, c(\omega_1)), \ldots, (\omega_m, c(\omega_m)))$: the labeled sample corresponding to S and regarded as drawn i.i.d. from $\mathscr{Z}$ according to the probability distribution $\tilde{P}$.
- $\hat{A}_{\tilde{S}}(g_h) = \frac{1}{m}\sum_{i=1}^m g_h(\omega_i, c(\omega_i)) = \frac{1}{m}\sum_{i=1}^m \Phi_\rho(\rho_{\tilde{h}}(\omega_i, c(\omega_i))) = \hat{R}_{S,\rho}(h)$, the empirical ρ-margin loss of h w.r.t. c on sample S.

- $\underset{z\sim\tilde{P}}{E}[g_h(z)] = \underset{\tilde{S}\sim\tilde{P}_m}{E}[\hat{A}_{\tilde{S}}(g_h)] = \underset{S\sim P_m}{E}[\hat{R}_{S,\rho}(h)] \geq \underset{S\sim P_m}{E}[\hat{R}_S(h)] = R(h)$.

By Theorem 3.1 of Lecture 2, for any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all g_h in $\mathcal{G}$:

$$
\begin{aligned}
\underset{z\sim\tilde{P}}{E}[g_h(z)] &\leq \frac{1}{m}\sum_{i=1}^{m} g_h(\omega_i, c(\omega_i)) + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \\
\underset{z\sim\tilde{P}}{E}[g_h(z)] &\leq \frac{1}{m}\sum_{i=1}^{m} g_h(\omega_i, c(\omega_i)) + 2\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}
\end{aligned}
$$

so that

$$R(h) \leq \hat{R}_{S,\rho}(h) + 2\mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

$$R(h) \leq \hat{R}_{S,\rho}(h) + 2\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) + 3\sqrt{\frac{\ln \frac{2}{\delta}}{2m}}$$

- The margin $\rho_{\tilde{h}}$ of the scoring function $\tilde{h}$, defined as

$$\rho_{\tilde{h}}(\omega, y) = \tilde{h}(\omega, y) - \max_{z \in \mathscr{Y}, z \neq y} \tilde{h}(\omega, z) \;\forall\, \omega \in \mathscr{I}, y \in \mathscr{Y},$$

  is a measurable function from $\mathscr{I} \times \mathscr{Y}$ to $\mathbb{R}$.

- $\mathcal{F} = \{\rho_{\tilde{h}} \mid h \in \mathcal{H}\}$.
  - Since $\sup_{h \in \mathcal{H}} |\tilde{h}(\omega, y)| < +\infty$ for all $\omega \in \mathscr{I}$ and for all $y \in \mathscr{Y}$, $\sup_{h \in \mathcal{H}} |\rho_{\tilde{h}}(\omega, y)| \leq \sum_{j \in \mathscr{Y}} \sup_{h \in \mathcal{H}} |\tilde{h}(\omega, j)| < +\infty$ for all $\omega \in \mathscr{I}$ and for all $y \in \mathscr{Y}$.

- $\mathcal{G} = \Phi_\rho \circ \mathcal{F}$.

- $\Phi_\rho : \mathbb{R} \to [0,1]$ is a $1/\rho$-Lipschitz function.
- Since $\sup_{h\in\mathcal{H}} |\rho_{\tilde{h}}(\omega, y)| < +\infty$ for all $\omega \in \mathscr{I}$ and for all $y \in \mathscr{Y}$, $\sup_{h\in\mathcal{H}} \left(\sum_{i=1}^{j} \sigma_i (\Phi_\rho \circ \rho_{\tilde{h}})(\omega_i, c(\omega_i)) + \sum_{i=j+1}^{m} \frac{1}{\rho} \sigma_i \rho_{\tilde{h}}(\omega_i, c(\omega_i))\right)$ is finite for all $\sigma_i \in \{-1, +1\}, i \in [1, m]$, for all $j \in [0, m]$ and for all labeled samples $S = (\omega_1, \ldots, \omega_m)$ of size m.

By Talagrand's lemma, we have

$$\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) \leq \frac{1}{\rho} \hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{F})$$

and then by taking expectation over $\tilde{S}$,

$$\mathfrak{R}_m(\mathcal{G}) \leq \frac{1}{\rho} \mathfrak{R}_m(\mathcal{F}).$$

Now

$$
\begin{aligned}
& \hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{F}) \\
= \quad & \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \rho_{\tilde{h}}(\omega_i, c(\omega_i)) \\
= \quad & \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \sum_{y \in \mathscr{Y}} \rho_{\tilde{h}}(\omega_i, y) 1_{y=c(\omega_i)} \\
\leq \quad & \sum_{y \in \mathscr{Y}} \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \rho_{\tilde{h}}(\omega_i, y) 1_{y=c(\omega_i)}
\end{aligned}
$$

by the sub-additivity of sup. But for each $y \in \mathscr{Y}$, we have

$$
\begin{aligned}
& \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \rho_{\tilde{h}}(\omega_i, y) 1_{y=c(\omega_i)} \\
= \quad & \underset{\sigma}{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \rho_{\tilde{h}}(\omega_i, y) \frac{(\epsilon_i + 1)}{2}\right], \ \epsilon_i \triangleq 2 1_{y=c(\omega_i)} - 1 \in \{-1, +1\} \\
\leq \quad & \frac{1}{2}\left(\underset{\sigma}{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \epsilon_i \rho_{\tilde{h}}(\omega_i, y)\right] + \underset{\sigma}{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \rho_{\tilde{h}}(\omega_i, y)\right]\right) \\
& \text{again by the sub-additivity of sup} \\
= \quad & \underset{\sigma}{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \rho_{\tilde{h}}(\omega_i, y)\right]
\end{aligned}
$$

so that

$$
\begin{aligned}
& \hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{F}) \\
\leq \quad & \sum_{y \in \mathscr{Y}} \underset{\sigma}{E} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \rho_{\tilde{h}}(\omega_i, y) \right] \\
= \quad & \sum_{y \in \mathscr{Y}} \underset{\sigma}{E} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\tilde{h}(\omega_i, y) - \max_{z \in \mathscr{Y}, z \neq y} \tilde{h}(\omega_i, z) \right) \right] \\
\leq \quad & \sum_{y \in \mathscr{Y}} \underset{\sigma}{E} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \tilde{h}(\omega_i, y) + \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} (-\sigma_i) \max_{z \in \mathscr{Y}, z \neq y} \tilde{h}(\omega_i, z) \right] \\
\leq \quad & \sum_{y \in \mathscr{Y}} \left(\underset{\sigma}{E} \left[\sup_{f \in \Pi_1(\mathcal{H})} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(\omega_i) \right] \right. \\
& \left. + \underset{\sigma}{E} \left[\sup_{f_j \in \Pi_1(\mathcal{H}), j \in [1, k-1]} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \max(f_1, f_2, \ldots, f_{k-1})(\omega_i) \right] \right).
\end{aligned}
$$

By Lemma 8.1, we have

$$\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{F}) \leq k^2 \underset{\sigma}{E}\left[\sup_{f \in \Pi_1(\mathcal{H})} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(\omega_i)\right] = k^2 \hat{\mathfrak{R}}_S(\Pi_1(\mathcal{H}))$$

and then

$$\hat{\mathfrak{R}}_{\tilde{S}}(\mathcal{G}) \leq \frac{k^2}{\rho}\hat{\mathfrak{R}}_S(\Pi_1(\mathcal{H}))$$

and then by taking expectation over $\tilde{S}$,

$$\mathfrak{R}_m(\mathcal{G}) \leq \frac{k^2}{\rho}\mathfrak{R}_m(\Pi_1(\mathcal{H})).$$

This completes the proof. $\square$

Remarks

- As other margin bounds presented in the previous chapters, the margin bounds in Theorem 8.1 show the trade-off between two terms: the larger the desired margin ρ, the smaller the middle term, at the price of a larger empirical multi-class classification margin loss $\hat{R}_{S,l}$.
- For the mono-label case of multi-class classification, there is additionally a quadratic dependency on the number k of classes. This suggests weaker guarantees when learning with a large number of classes or the need for even larger margins ρ for which the empirical margin loss would be small.
- We will derive a simple upper bound for the Rademacher complexity of $\Pi_1(\mathcal{H})$ for kernel-based hypotheses.

Kernel-Based Hypotheses for Multi-Class Classification

- $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a PDS kernel over the input space $\mathscr{I}$.
- $\Phi : \mathscr{I} \to \mathscr{F}$: a feature mapping associated to the PDS kernel K from the input space $\mathscr{I}$ to the feature space $\mathscr{F}$, which is a Hilbert space, so that

$$K(\omega, \omega') = \langle \Phi(\omega), \Phi(\omega') \rangle_{\mathscr{F}}.$$

- In multi-class classification, a kernel-based hypothesis is based on k weight vectors $f_1, \ldots, f_k$ in the feature space $\mathscr{F}$.

- Each weight vector $f_y, y \in [1,k]$, defines a scoring function

$$\omega \to \langle f_y, \Phi(\omega) \rangle_{\mathscr{F}}$$

and the predicted class of the item $\omega \in \mathscr{I}$ is given by

$$\arg\max_{y \in \mathscr{Y}} \langle f_y, \Phi(\omega) \rangle_{\mathscr{F}}.$$

  – If the feature space $\mathscr{F}$ is the RKHS $\mathbb{H}$ of K, the reproducing property gives

$$\langle f_y, \Phi(\omega) \rangle_{\mathbb{H}} = f_y(\omega).$$

- $\mathbf{f} = [f_1, \ldots, f_k]^T$: the vector formed by the k weight vectors $f_y, y \in [1,k]$, in the feature space $\mathscr{F}$.
- $\|\mathbf{f}\|_{\mathscr{F},p} = \left(\sum_{i=1}^k \|f_i\|_{\mathscr{F}}^p \right)^{1/p}$: the $L_{\mathscr{F},p}$-norm of $\mathbf{f}$, where $p \geq 1$.

- $\mathcal{H}_{K,\mathscr{F},p,\Lambda} = \{\omega \mapsto \arg\max_{y\in\mathscr{Y}} \langle f_y, \Phi(\omega)\rangle_{\mathscr{F}} \mid \|\mathbf{f}\|_{\mathscr{F},p} \leq \Lambda$, where $\mathbf{f} = [f_1, \ldots, f_k]^T\}$: the kernel-based hypothesis set we will consider.
  - The scoring function corresponding to a hypothesis $h \in \mathcal{H}_{K,\mathscr{F},p,\Lambda}$ is

$$\begin{aligned} \tilde{h}(\omega, y) &= \langle f_y, \Phi(\omega)\rangle_{\mathscr{F}} \\ &= f_y(\omega) \text{ if } \mathscr{F} \text{ is the RKHS } \mathbb{H} \text{ of } K. \end{aligned}$$

Rademacher Complexity of Multi-Class Kernel-Based Hypotheses

Proposition 8.1: Let

- $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a PDS kernel over the input space $\mathscr{I}$.
- $\Phi : \mathscr{I} \to \mathscr{F}$: a feature mapping associated to the PDS kernel K from the input space $\mathscr{I}$ to the feature space $\mathscr{F}$ so that $K(\omega, \omega') = \langle \Phi(\omega), \Phi(\omega) \rangle_{\mathscr{F}}$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of size m.

Assume that

- there is an $r > 0$ such that $K(\omega, \omega) \leq r^2$ for all $\omega \in \mathscr{I}$.

Then for any $m \geq 1$, we have

$$\hat{\mathfrak{R}}_S(\Pi_1(\mathcal{H}_{K,\mathscr{F},p,\Lambda})) \leq \sqrt{\frac{r^2 \Lambda^2}{m}}.$$

Proof.

- The condition $\|\mathbf{f}\|_{\mathscr{F},p} \leq \Lambda$ implies that $\|f_y\|_{\mathscr{F}} \leq \Lambda$ for all $y \in [1, k]$ since

$$\|f_y\|_{\mathscr{F}} = \left(\|f_y\|_{\mathscr{F}}^p\right)^{1/p} \leq \left(\sum_{z=1}^{k} \|f_z\|_{\mathscr{F}}^p\right)^{1/p} = \|\mathbf{f}\|_{\mathscr{F},p} \leq \Lambda.$$

Now

$$
\begin{aligned}
& \hat{\mathfrak{R}}_S(\Pi_1(\mathcal{H}_{K,\mathscr{F},p,\Lambda})) \\
= \quad & \underset{\sigma}{E}\left[\sup_{f\in\Pi_1(\mathcal{H}_{K,\mathscr{F},p,\Lambda})} \frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\omega_i)\right] \\
\leq \quad & \frac{1}{m}\underset{\sigma}{E}\left[\sup_{\|f\|_{\mathscr{F}}\leq\Lambda} \sum_{i=1}^{m}\sigma_i\langle f,\Phi(\omega_i)\rangle_{\mathscr{F}}\right] \\
= \quad & \frac{1}{m}\underset{\sigma}{E}\left[\sup_{\|f\|_{\mathscr{F}}\leq\Lambda} \langle f,\sum_{i=1}^{m}\sigma_i\Phi(\omega_i)\rangle_{\mathscr{F}}\right] \\
\leq \quad & \frac{1}{m}\underset{\sigma}{E}\left[\sup_{\|f\|_{\mathscr{F}}\leq\Lambda} \|f\|_{\mathscr{F}}\left\|\sum_{i=1}^{m}\sigma_i\Phi(\omega_i)\right\|_{\mathscr{F}}\right] \\
& \text{by Cauchy-Schwartz inequality} \\
= \quad & \frac{\Lambda}{m}\underset{\sigma}{E}\left[\left\|\sum_{i=1}^{m}\sigma_i\Phi(\omega_i)\right\|_{\mathscr{F}}\right].
\end{aligned}
$$

Since $x \mapsto \sqrt{x}$ is concave for all $x \geq 0$, by Jensen's inequality, we have

$$\begin{aligned}
& \underset{\sigma}{E}\left[\left\|\sum_{i=1}^{m} \sigma_i \Phi(\omega_i)\right\|_{\mathscr{F}}\right] = \underset{\sigma}{E}\left[\sqrt{\left\|\sum_{i=1}^{m} \sigma_i \Phi(\omega_i)\right\|_{\mathscr{F}}^2}\right] \\
\leq & \sqrt{\underset{\sigma}{E}\left[\left\|\sum_{i=1}^{m} \sigma_i \Phi(\omega_i)\right\|_{\mathscr{F}}^2\right]} = \sqrt{\underset{\sigma}{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{m} \sigma_i \sigma_j \langle \Phi(\omega_i), \Phi(\omega_j)\rangle_{\mathscr{F}}\right]} \\
= & \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} \underset{\sigma}{E}\left[\sigma_i \sigma_j\right] \langle \Phi(\omega_i), \Phi(\omega_j)\rangle_{\mathscr{F}}} \\
= & \sqrt{\sum_{i=1}^{m} \langle \Phi(\omega_i), \Phi(\omega_i)\rangle_{\mathscr{F}}} = \sqrt{\sum_{i=1}^{m} K(\omega_i, \omega_i)} \quad \text{since } \underset{\sigma}{E}\left[\sigma_i \sigma_j\right] = \delta_{ij} \\
\leq & \sqrt{mr^2}.
\end{aligned}$$

We conclude that

$$\hat{\mathfrak{R}}_S(\Pi_1(\mathcal{H}_{K,\mathscr{F},p,\Lambda})) \leq \frac{\Lambda\sqrt{mr^2}}{m} = \sqrt{\frac{\Lambda^2 r^2}{m}}.$$

$\square$

Margin Bound for Multi-Class Classification with Kernel-Based Hypotheses

Corollary 8.1: Let

- $\mathscr{I}$: the input space of all possible items ω, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$.
- $c : \mathscr{I} \to \mathscr{Y} = \{1, 2, \ldots, k\}$: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a PDS kernel over the input space $\mathscr{I}$.
- $\Phi : \mathscr{I} \to \mathscr{F}$: a feature mapping associated to the PDS kernel K from the input space $\mathscr{I}$ to the feature space $\mathscr{F}$ so that $K(\omega, \omega') = \langle \Phi(\omega), \Phi(\omega') \rangle_{\mathscr{F}}$.
- $S = (\omega_1, \ldots, \omega_m)$: a labeled sample of size m with labels $(c(\omega_1), \ldots, c(\omega_m))$.

- $\rho > 0$: a given margin.
- $\hat{R}_{S,\rho}(h) \triangleq \frac{1}{m}\sum_{i=1}^{m} \Phi_\rho(\rho_{\tilde{h}}(\omega_i, c(\omega_i)))$: the empirical ρ-margin loss of the hypothesis h for multi-class classification w.r.t. the concept c on a labeled sample S of size m.
- $p \geq 1$.
- $\mathcal{H}_{K,\mathscr{F},p,\Lambda} = \{\omega \mapsto \arg\max_{y\in\mathscr{Y}} \langle f_y, \Phi(\omega)\rangle_{\mathscr{F}} \mid \|\mathbf{f}\|_{\mathscr{F},p} \leq \Lambda$, where $\mathbf{f} = [f_1, \ldots, f_k]^T\}$: the kernel-based hypothesis set.

Assume that

- there is an $r > 0$ such that $K(\omega, \omega) \leq r^2$ for all $\omega \in \mathscr{I}$.

Then for any $\delta > 0$, with probability at least $1 - \delta$, the following multi-class classification generalization bound holds for all h in $\mathcal{H}_{K,\mathscr{F},p,\Lambda}$:

$$R(h) \quad \leq \quad \hat{R}_{S,\rho}(h) + 2k^2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}.$$

Proof. We first justify that

$$
\begin{aligned}
\sup_{h \in \mathcal{H}_{K,\mathscr{F},p,\Lambda}} |\tilde{h}(\omega, y)| &\leq \sup_{\|f\|_{\mathscr{F}} \leq \Lambda} |\langle f, \Phi(\omega)\rangle_{\mathscr{F}}| \\
&\leq \sup_{\|f\|_{\mathscr{F}} \leq \Lambda} \|f\|_{\mathscr{F}} \|\Phi(\omega)\|_{\mathscr{F}} \text{ by Cauchy-Schwartz inequality} \\
&\leq \Lambda \|\Phi(\omega)\|_{\mathscr{F}} < +\infty
\end{aligned}
$$

for all $\omega \in \mathscr{I}$ and for all $y \in \mathscr{Y}$. The corollary now follows from Theorem 8.1 and Proposition 8.1. $\square$

Two Families of Multi-Class Classification Algorithms

- Single classifier:
  - Multi-class SVMs.
  - AdaBoost.MH.
  - Decision trees : often used as base classifiers in boosting.
- Combination of binary classifiers: reducing the problem of multi-class classification to that of multiple binary classification tasks, training a binary classification algorithm for each of these tasks independently and defining the multi-class predictor as a combination of the hypotheses returned by each of these algorithms.
  - One-vs-all.
  - One-vs-one.
  - Error-correcting codes.

The Contents of This Lecture

- Multi-class classification problem.
- Generalization bound.
- Uncombined multi-class algorithms.
- Aggregated multi-class algorithms.

A Generalization Guarantee of Kernel-Based Hypotheses for Multi-Class Classification

- $\mathscr{I}$: the input space of all possible items ω, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$.
- $c : \mathscr{I} \to \mathscr{Y} = \{1, 2, \ldots, k\}$: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from $\mathscr{I}$ according to the distribution P with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a PDS kernel over the input space $\mathscr{I}$.
- $\Phi : \mathscr{I} \to \mathscr{F}$: a feature mapping associated to the PDS kernel K from the input space $\mathscr{I}$ to the feature space $\mathscr{F}$ so that $K(\omega, \omega') = \langle \Phi(\omega), \Phi(\omega') \rangle_{\mathscr{F}}$.

- $\mathcal{H}_{K,\mathscr{F},2,\Lambda} = \{\omega \mapsto \arg\max_{y\in\mathscr{Y}} \langle f_y, \Phi(\omega)\rangle_{\mathscr{F}} \mid \|\mathbf{f}\|_{\mathscr{F},2} = \sqrt{\sum_{y\in\mathscr{Y}} \|f_y\|^2_{\mathscr{F}}} \leq \Lambda$, where $\mathbf{f} = [f_1, \ldots, f_k]^T\}$: the kernel-based hypothesis set.
  - Scoring functions are $\tilde{h}(\omega, y) = \langle f_y, \Phi(\omega)\rangle_{\mathscr{F}}$.
  - The margin function of the scoring function $\tilde{h}$ is

$$\begin{aligned} \rho_{\tilde{h}}(\omega, y) &= \tilde{h}(\omega, y) - \max_{z\in\mathscr{Y}, z\neq y} \tilde{h}(\omega, z) \\ &= \langle f_y, \Phi(\omega)\rangle_{\mathscr{F}} - \max_{z\in\mathscr{Y}, z\neq y} \langle f_z, \Phi(\omega)\rangle_{\mathscr{F}}. \end{aligned}$$

- The 1-margin loss function $\Phi_1(y')$ is no more than the hinge loss function $\max(0, 1-y')$ for all $y' \in \mathbb{R}$ so that

$$\Phi_1(\rho_{\tilde{h}}(\omega, y)) \leq \max(0, 1-\rho_{\tilde{h}}(\omega, y)),$$

that is,

$$\begin{aligned} & \Phi_1\left(\langle f_y, \Phi(\omega)\rangle_{\mathscr{F}} - \max_{z\in\mathscr{Y}, z\neq y} \langle f_z, \Phi(\omega)\rangle_{\mathscr{F}}\right) \\ \leq \quad & \max\left(0, 1 - \langle f_y, \Phi(\omega)\rangle_{\mathscr{F}} + \max_{z\in\mathscr{Y}, z\neq y} \langle f_z, \Phi(\omega)\rangle_{\mathscr{F}}\right). \end{aligned}$$

  – The 1-margin loss function $\Phi_1(y')$ is neither convex nor concave; but the hinge loss function $\max(0, 1-y')$ is convex.

- $\hat{R}_{S,1}(h) \triangleq \frac{1}{m}\sum_{i=1}^{m}\Phi_1(\rho_{\tilde{h}}(\omega_i, c(\omega_i)))$: the empirical 1-margin loss of the hypothesis h for multi-class classification w.r.t. the concept c on a labeled sample S of size m, which is upper bounded by

$$\hat{R}_{S,1}(h) \leq \frac{1}{m}\sum_{i=1}^{m}\eta_i,$$

where η_i is the slack variable which compensates the deficit of the margin $\tilde{h}(\omega_i, c(\omega_i))$ of the i-th item ω_i of the random sample S from 1:

$$\eta_i \triangleq \max\left(0, 1 - \langle f_{c(\omega_i)}, \Phi(\omega_i)\rangle_{\mathscr{F}} + \max_{z\in\mathscr{Y}, z\neq c(\omega_i)}\langle f_z, \Phi(\omega_i)\rangle_{\mathscr{F}}\right).$$

Assume that

- there is an $r > 0$ such that $K(\omega, \omega) \leq r^2$ for all $\omega \in \mathscr{I}$.

Then Corollary 8.1 shows that for any $\delta > 0$, with probability at least $1 - \delta$, the following multi-class classification generalization bound holds for all h in $\mathcal{H}_{K,\mathscr{F},2,\Lambda}$:

$$\begin{aligned} R(h) &\leq \hat{R}_{S,1}(h) + 2k^2\sqrt{\frac{r^2\Lambda^2}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \\ &\leq \frac{1}{m}\sum_{i=1}^{m} \eta_i + 2k^2\sqrt{\frac{r^2\Lambda^2}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}. \end{aligned}$$

- To minimize the upper bound of the generalization guarantee, we have to minimize $\sum_{i=1}^{m} \eta_i$ and $\sum_{y \in \mathscr{Y}} \|f_y\|_{\mathscr{F}}^2$ (which can be set to Λ^2) simultaneously.

Multi-Class Kernel-Based SVMs

- Optimization problem:

$$\begin{aligned}
\text{Minimize} \quad & F(\mathbf{f}, \eta) = \frac{1}{2} \sum_{y=1}^{k} \|f_y\|_{\mathscr{F}}^2 + C \sum_{i=1}^{m} \eta_i \\
\text{Subject to} \quad & 1 - \eta_i - \langle f_{c(\omega_i)}, \Phi(\omega_i) \rangle_{\mathscr{F}} + \langle f_y, \Phi(\omega_i) \rangle_{\mathscr{F}} \leq 0, \\
& \qquad \forall\ i \in [1, m] \text{ and } y \in [1, k], y \neq c(\omega_i) \\
& -\eta_i \leq 0,\ \forall\ i \in [1, m] \\
& (\mathbf{f}, \eta) \in \mathscr{F}^k \times \mathbb{R}^m.
\end{aligned}$$

  - If the feature space $\mathscr{F}$ is the RKHS $\mathbb{H}$ of the kernel K, we have $\langle f_y, \Phi(\omega_i) \rangle_{\mathbb{H}} = f_y(\omega)$.
  - We will employ the RKHS $\mathbb{H}$ of the kernel K as the feature space.

Remarks

- The parameter $C > 0$ determines the trade-off between margin-maximization (or minimization of $\sum_{y=1}^{k} \|f_y\|_{\mathscr{F}}^2$) and the minimization of the slack penalty $\sum_{i=1}^{m} \eta_i$.
- The parameter C is typically determined via n-fold cross-validation.

The Primal Problem for Multi-Class Kernel-Based SVM

$$
\begin{array}{ll}
\text{Minimize} & F(\mathbf{f}, \eta) = \frac{1}{2}\sum_{y=1}^{k} \|f_y\|_{\mathbb{H}}^2 + C\sum_{i=1}^{m} \eta_i \\
\text{Subject to} & 1 - \eta_i - f_{c(\omega_i)}(\omega_i) + f_y(\omega_i) \leq 0, \\
 & \qquad \forall\ i \in [1, m] \text{ and } y \in [1, k], y \neq c(\omega_i) \\
 & -\eta_i \leq 0,\ \forall\ i \in [1, m] \\
 & (\mathbf{f}, \eta) \in \mathbb{H}^k \times \mathbb{R}^m.
\end{array}
$$

- How do we solve this primal problem when the RKHS $\mathscr{F}$ of the kernel K is an infinite-dimensional Hilbert space ?
- We need a generalization of the Representer Theorem in Lecture 4.

A Generalization of the Representer Theorem

Let

- $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a PDS kernel over an input space $\mathscr{I}$.
- $\mathscr{F} = \mathbb{H}$: a feature space, which is the reproducing kernel Hilbert space (RKHS) associated to the PDS kernel K.
- $(\omega_1, \omega_2, \ldots, \omega_m)$: a given m-tuple over the input space $\mathscr{I}$.
- $G : (\mathbb{R}^+)^k \to \mathbb{R}$: a non-decreasing function in each of the k arguments.
- $L : \mathbb{R}^{km} \to \mathbb{R} \cup \{\infty\}$: any function.

Any solution of the optimization problem

$$\text{Minimize}_{f_1,\ldots,f_k\in\mathbb{H}}\ F(f_1,\ldots,f_k) = G(\|f_1\|_{\mathbb{H}},\ldots,\|f_k\|_{\mathbb{H}})$$
$$+L(f_1(\omega_1),\ldots,f_k(\omega_1),\ldots,f_1(\omega_m),\ldots,f_k(\omega_m))$$

admits a solution of the form

$$f_j^* = \sum_{i=1}^{m}\alpha_{i,j}K(\omega_i,\cdot),\ \ j\in[1,k]$$

for some real numbers $\alpha_{i,j}, i\in[1,m], j\in[1,k]$. If G is further assumed to be strictly increasing in each of the k arguments, then any solution has this form.

Proof.

- $\mathbb{H}_1 = \text{Span}(\{K(\omega_i,\cdot), i\in[1,m]\})$: a finite-dimensional subspace of the RKHS $\mathbb{H}$, which is a closed subspace.
  - Closedness: if a sequence $\{h_n\}_{n=1}^{\infty}$ in $\mathbb{H}_1$ converges to an $h\in\mathbb{H}$, then h must be in $\mathbb{H}_1$.

- $\mathbb{H}_1^{\perp} = \{h \in \mathbb{H} : \langle h, h' \rangle = 0 \;\forall\; h' \in \mathbb{H}_1\}$: the orthogonal complement of $\mathbb{H}_1$, which is a closed subspace of $\mathbb{H}$.
- Since $\mathbb{H}_1$ is closed, $\mathbb{H} = \mathbb{H}_1 \oplus \mathbb{H}_1^{\perp}$, i.e., $\mathbb{H}$ is the direct sum of $\mathbb{H}_1$ and $\mathbb{H}_1^{\perp}$, which means that for each $f_j \in \mathbb{H}$, there exist unique $h_j \in \mathbb{H}_1$ and $h_j^{\perp} \in \mathbb{H}_1^{\perp}$ such that $f_j = h_j + h_j^{\perp}$.
- Since G is non-decreasing in each of the k arguments, we have

$$
\begin{aligned}
& G(\|h_1\|_{\mathbb{H}}, \|h_2\|_{\mathbb{H}}, \ldots, \|h_k\|_{\mathbb{H}}) \\
\leq \;& G(\|f_1\|_{\mathbb{H}}, \|h_2\|_{\mathbb{H}}, \ldots, \|h_k\|_{\mathbb{H}}) \text{ since} \|h_1\|_{\mathbb{H}} \leq \|f_1\|_{\mathbb{H}} \\
\vdots \; & \\
\leq \;& G(\|f_1\|_{\mathbb{H}}, \|f_2\|_{\mathbb{H}}, \ldots, \|f_k\|_{\mathbb{H}}).
\end{aligned}
$$

- By the reproducing property, for all $i \in [1, m]$, $j \in [1, k]$, $f_j(\omega_i) = \langle f_j, K(\omega_i, \cdot)\rangle = \langle h_j, K(\omega_i, \cdot)\rangle = h_j(\omega_i)$. Thus, $L(f_1(\omega_1), \ldots, f_k(\omega_1), \ldots, f_1(\omega_m), \ldots, f_k(\omega_m)) = L(h_1(\omega_1), \ldots, h_k(\omega_1), \ldots, h_1(\omega_m), \ldots, h_k(\omega_m))$.

- $F(h_1, \ldots, h_k) \leq F(f_1, \ldots, f_k)$ for all $f_1, \ldots, f_k \in \mathbb{H}$, which proves the first part of the theorem.

- If G is further strictly increasing, then $F(f_1, \ldots, h_j, \ldots, f_k) < F(f_1, \ldots, f_j, \ldots, f_k)$ when $\|h_j^{\perp}\|_{\mathbb{H}} > 0$ and any solution of the optimization problem must be in $\mathbb{H}_1^k$.

$\square$

Reformulation of Primal Problem for Multi-Class Kernel-Based SVM

- $\mathscr{I}$: the input space of all possible items, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$, where P is unknown.
- $c : \mathscr{I} \to \mathscr{Y} = \{1, 2, \ldots, k\}$: a fixed but unknown concept.
- $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a PDS kernel over the input space $\mathscr{I}$.
- $\mathscr{F} = \mathbb{H}$: a feature space, which is the reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ associated to the PDS kernel K with the feature mapping $\Phi : \mathscr{I} \to \mathbb{H}$ such that $\Phi(\omega) = K(\omega, \cdot)$.
- $S = (\omega_1, \omega_2, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to the distribution P with labels $(c(\omega_1), c(\omega_2), \ldots, c(\omega_m))$.

The primal problem for multi-class SVM in the RKHS feature space $\mathbb{H}$ associated to the PDS kernel K is

$$
\begin{aligned}
&\text{Minimize} && F(\mathbf{f}, \eta) = \tfrac{1}{2} \textstyle\sum_{y=1}^{k} \|f_y\|_{\mathbb{H}}^2 + C \sum_{i=1}^{m} \eta_i \\
&\text{Subject to} && 1 - \eta_i - f_{c(\omega_i)}(\omega_i) + f_y(\omega_i) \leq 0, \\
& && \qquad\qquad \forall\ i \in [1, m] \text{ and } y \in [1, k], y \neq c(\omega_i) \\
& && -\eta_i \leq 0,\ \forall\ i \in [1, m] \\
& && (\mathbf{f}, \eta) \in \mathbb{H}^k \times \mathbb{R}^m.
\end{aligned}
$$

which is equivalent to

$$
\text{Minimize}_{f_1, \ldots, f_k \in \mathbb{H}}\ \tilde{F}(f_1, \ldots, f_k) = \frac{1}{2} \sum_{y=1}^{k} \|f_y\|_{\mathbb{H}}^2
$$

$$
+C \sum_{i=1}^{m} \max(0, 1 - f_{c(\omega_i)}(\omega_i) + \max_{y \in [1,k], y \neq c(\omega_i)} f_y(\omega_i)).
$$

By letting

- $G(\|f_1\|_{\mathbb{H}}, \ldots, \|f_k\|_{\mathbb{H}}) = \frac{1}{2}\sum_{j=1}^{k} \|f_j\|_{\mathbb{H}}^2$ with $G(x_1, \ldots, x_k) = \frac{1}{2}\sum_{j=1}^{k} x_j^2$ strictly increasing in each of the k arguments;
- $L(f_1(\omega_1), \ldots, f_k(\omega_1), \ldots, f_1(\omega_m), \ldots, f_k(\omega_m)) = C\sum_{i=1}^{m} \max(0, 1 - f_{c(\omega_i)}(\omega_i) + \max_{y\in[1,k], y\neq c(\omega_i)} f_y(\omega_i))$,

any solution of the optimization problem

$$\text{Minimize}_{f_1,\ldots,f_k\in\mathbb{H}}\ \tilde{F}(f_1, \ldots, f_k) = \frac{1}{2}\sum_{y=1}^{k} \|f_y\|_{\mathbb{H}}^2$$
$$+C\sum_{i=1}^{m} \max(0, 1 - f_{c(\omega_i)}(\omega_i) + \max_{y\in[1,k], y\neq c(\omega_i)} f_y(\omega_i)).$$

must be of the form $f_j^* = \sum_{i=1}^{m} \alpha_{i,j} K(\omega_i, \cdot)$ by the generalization of the representer theorem.

Let

$$\begin{aligned}
\mathbb{H}_S &\triangleq \operatorname{Span}\{K(\omega_i,\cdot), i=1,2,\ldots,m\} \\
&= \left\{\sum_{i=1}^{m} \beta_i K(\omega_i,\cdot) \mid \beta_i \in \mathbb{R},\ 1 \le i \le m\right\},
\end{aligned}$$

which is a finite-dimensional Hilbert space. Then we have

$$\begin{aligned}
& \text{Minimize}_{f_1,\ldots,f_k \in \mathbb{H}}\ \tilde{F}(f_1,\ldots,f_k) = \frac{1}{2}\sum_{y=1}^{k} \|f_y\|_{\mathbb{H}}^2 \\
& \qquad + C\sum_{i=1}^{m} \max(0, 1 - f_{c(\omega_i)}(\omega_i) + \max_{y\in[1,k], y\neq c(\omega_i)} f_y(\omega_i)) \\
\Leftrightarrow\ & \text{Minimize}_{f_1,\ldots,f_k \in \mathbb{H}_S}\ \tilde{F}(f_1,\ldots,f_k) = \frac{1}{2}\sum_{y=1}^{k} \|f_y\|_{\mathbb{H}_S}^2 \\
& \qquad + C\sum_{i=1}^{m} \max(0, 1 - f_{c(\omega_i)}(\omega_i) + \max_{y\in[1,k], y\neq c(\omega_i)} f_y(\omega_i)).
\end{aligned}$$

Thus the primal problem for multi-class SVM in the RKHS feature space $\mathbb{H}$ associated to the PDS kernel K is equivalent to

$$
\begin{aligned}
\text{Minimize} \quad & F(\mathbf{f}, \eta) = \tfrac{1}{2} \textstyle\sum_{y=1}^{k} \|f_y\|_{\mathbb{H}_S}^2 + C \textstyle\sum_{i=1}^{m} \eta_i \\
\text{Subject to} \quad & 1 - \eta_i - f_{c(\omega_i)}(\omega_i) + f_y(\omega_i) \leq 0, \\
& \qquad\qquad \forall\ i \in [1, m] \text{ and } y \in [1, k], y \neq c(\omega_i) \\
& -\eta_i \leq 0,\ \forall\ i \in [1, m] \\
& (\mathbf{f}, \eta) \in \mathbb{H}_S^k \times \mathbb{R}^m.
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
\text{Minimize} \quad & F(\mathbf{f}, \eta) = \tfrac{1}{2} \textstyle\sum_{y=1}^{k} \|f_y\|_{\mathbb{H}_S}^2 + C \textstyle\sum_{i=1}^{m} \eta_i \\
\text{Subject to} \quad & 1 - \eta_i - \langle f_{c(\omega_i)}, \Phi(\omega_i) \rangle + \langle f_y, \Phi(\omega_i) \rangle \leq 0, \\
& \qquad\qquad \forall\ i \in [1, m] \text{ and } y \in [1, k], y \neq c(\omega_i) \\
& -\eta_i \leq 0,\ \forall\ i \in [1, m] \\
& (\mathbf{f}, \eta) \in \mathbb{H}_S^k \times \mathbb{R}^m.
\end{aligned}
$$

Qualification of the Primal Problem for Multi-Class Kernel-Based SVM

- The object function $F(\mathbf{f}, \eta) = \frac{1}{2}\sum_{y=1}^{k} \|f_y\|_{\mathbb{H}_S}^2 + C\sum_{i=1}^{m} \eta_i$ is infinitely differentiable and convex so that it is pseudoconvex at any feasible point.
  - Each f_y in $\mathbb{H}_S$ is regarded as the coordinate vector with respective to a basis of the finite-dimensional Hilbert space $\mathbb{H}_S$.
- The inequality constraint functions $g_{iy}(\mathbf{f}, \eta) = 1 - \eta_i - \langle f_{c(\omega_i)}, \Phi(\omega_i)\rangle_{\mathbb{H}_S} + \langle f_y, \Phi(\omega_i)\rangle_{\mathbb{H}_S}$, $i \in [1, m]$ and $y \in [1, k], y \neq c(\omega_i)$ and $h_i(\mathbf{f}, \eta) = -\eta_i$, $i \in [1, m]$, are affine functions so that they are infinitely differentiable and convex and then quasiconvex at any feasible point.

- $\nabla F = \begin{bmatrix} f_1 \\ \vdots \\ f_k \\ C\mathbf{1}^{(m)} \end{bmatrix}$, $\nabla g_{iy} = \begin{bmatrix} \Phi(\omega_i) \otimes (-\mathbf{e}^{(k)}_{c(\omega_i)} + \mathbf{e}^{(k)}_y) \\ -\mathbf{e}^{(m)}_i \end{bmatrix}$, and

  $\nabla h_i = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ -\mathbf{e}^{(m)}_i \end{bmatrix}$, where $\mathbf{e}^{(k)}_i$ and $\mathbf{e}^{(m)}_i$ are standard unit vectors in $\mathbb{R}^k$ and $\mathbb{R}^m$ respectively and the Kronecker product $\Phi(\omega_i) \otimes \mathbf{e}^{(k)}_y$ means $[0, \ldots, 0, \underbrace{\Phi(\omega_i)^T}_{y\text{th position}}, 0, \ldots, 0]^T$.

- The Kuhn-Tucker necessary conditions are:

$$\nabla F + \sum_{i=1}^{m} \sum_{y=1, y \neq c(\omega_i)}^{k} \lambda_{iy} \nabla g_{iy} + \sum_{i=1}^{m} \mu_i \nabla h_i = \mathbf{0}$$
$$\iff f_y = \sum_{i=1, c(\omega_i)=y}^{m} \sum_{z=1, z \neq y}^{k} \lambda_{iz} \Phi(\omega_i)$$
$$- \sum_{i=1, c(\omega_i) \neq y}^{m} \lambda_{iy} \Phi(\omega_i), y \in [1, k]$$
$$\text{and } C = \mu_i + \sum_{y=1, y \neq c(\omega_i)}^{k} \lambda_{iy}, i \in [1, m],$$
$$\lambda_{iy} g_{iy}(\mathbf{f}, \eta) = 0, \; i \in [1, m], y \in [1, k], y \neq c(\omega_i),$$
$$\mu_i \eta_i = 0, \; i \in [1, m],$$
$$\lambda_{iy} \geq 0, \; i \in [1, m], y \in [1, k], y \neq c(\omega_i),$$
$$\mu_i \geq 0, \; i \in [1, m].$$

- Any feasible point $(\mathbf{f}, \eta)$ which satisfies the Kuhn-Tucker necessary conditions in above is a global minimum solution.

Support Vectors

- Support vectors for class $y, y \in \mathscr{Y}$: any vector $\Phi(\omega_i)$ which appears in the linear combination

$$f_y = \sum_{i=1, c(\omega_i)=y}^{m} \sum_{z=1, z \neq y}^{k} \lambda_{iz} \Phi(\omega_i) - \sum_{i=1, c(\omega_i) \neq y}^{m} \lambda_{iy} \Phi(\omega_i),$$

  i.e., $\sum_{z=1, z\neq y}^{k} \lambda_{iz} \neq 0$ for those i such that $c(\omega_i) = y$ and $\lambda_{iy} \neq 0$ for those i such that $c(\omega_i) \neq y$.

- If $\lambda_{iy} \neq 0$, we must have $\langle f_{c(\omega_i)} - f_y, \Phi(\omega_i) \rangle_{\mathbb{H}_S} = 1 - \eta_i$ by the complementary slackness conditions.
  - Furthermore, if $\eta_i = 0$, then $\langle f_{c(\omega_i)} - f_y, \Phi(\omega_i) \rangle_{\mathbb{H}_S} = 1$.
- If $\eta_i > 0$, then $\mu_i = 0$ and then $\sum_{z=1, z \neq c(\omega_i)}^{k} \lambda_{iz} = C > 0$ so that $\Phi(\omega_i)$ is a support vector of the weight vector $f_{c(\omega_i)}$. This $\Phi(\omega_i)$ is called an outlier w.r.t. the weight vector $f_{c(\omega_i)}$.

How to Determine Optimal Lagrangian Variables λ_{iy}^{SVM} ?

- Once optimal Lagrangian variables λ_{iy}^{SVM} are determined, we can compute

$$f_y^{SVM} = \sum_{i=1, c(\omega_i)=y}^{m} \sum_{z=1, z \neq y}^{k} \lambda_{iz}^{SVM} \Phi(\omega_i) - \sum_{i=1, c(\omega_i) \neq y}^{m} \lambda_{iy}^{SVM} \Phi(\omega_i).$$

- We will use the Lagrangian dual problem to determine optimal λ_{iy}^{SVM}.

Lagrangian Dual Function for Multi-Class Kernel-Based SVM

- $X = \mathbb{H}_S^k \times \mathbb{R}^m$: a nonempty open convex set.
- Lagrangian function: for all $\mathbf{f} \in \mathbb{H}_S^k, \eta \in \mathbb{R}^m$, and $\lambda \in \mathbb{R}^{m(k-1)}, \mu \in \mathbb{R}^m$,

$$\begin{aligned}
& L(\mathbf{f}, \eta, \lambda, \mu) \\
= \;& F(\mathbf{f}, \eta) + \sum_{i=1}^{m} \sum_{y=1, y \neq c(\omega_i)}^{k} \lambda_{iy} g_{iy}(\mathbf{f}, \eta) + \sum_{i=1}^{m} \mu_i h_i(\mathbf{f}, \eta) \\
= \;& \frac{1}{2} \sum_{y=1}^{k} \|f_y\|^2 + C \sum_{i=1}^{m} \eta_i + \sum_{i=1}^{m} \sum_{y=1, y \neq c(\omega_i)}^{k} \\
& \lambda_{iy} (1 - \eta_i - \langle f_{c(\omega_i)} - f_y, \Phi(\omega_i) \rangle_{\mathbb{H}_S}) - \sum_{i=1}^{m} \mu_i \eta_i .
\end{aligned}$$

- For any fixed $\lambda \in \mathbb{R}^{m(k-1)}, \mu \in \mathbb{R}^m$, the gradient ∇L of the Lagrangian function w.r.t. $(\mathbf{f}, \eta)$ is

$$
\begin{aligned}
\nabla L &= \nabla F + \sum_{i=1}^{m} \sum_{y=1, y \neq c(\omega_i)}^{k} \lambda_{iy} \nabla g_{iy} + \sum_{i=1}^{m} \mu_i \nabla h_i \\
&= \begin{bmatrix} f_1 \\ \vdots \\ f_k \\ C\mathbf{1}^{(m)} \end{bmatrix} + \sum_{i=1}^{m} \sum_{y=1, y \neq c(\omega_i)}^{k} \lambda_{iy} \begin{bmatrix} \Phi(\omega_i) \otimes (-\mathbf{e}_{c(\omega)}^{(k)} + \mathbf{e}_y^{(k)}) \\ -\mathbf{e}_i^{(m)} \end{bmatrix} \\
&\quad + \sum_{i=1}^{m} \mu_i \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ -\mathbf{e}_i^{(m)} \end{bmatrix}
\end{aligned}
$$

and the Hessian matrix is

$$\mathbf{H} = \begin{bmatrix} I_{\mathbb{H}_S} & \cdots & \mathbf{0}_{\mathbb{H}_S} & \mathbf{0}_{\dim(\mathbb{H}_S)\times m} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_{\mathbb{H}_S} & \cdots & I_{\mathbb{H}_S} & \mathbf{0}_{\dim(\mathbb{H}_S)\times m} \\ \mathbf{0}_{m\times\dim(\mathbb{H}_S)} & \cdots & \mathbf{0}_{m\times\dim(\mathbb{H}_S)} & \mathbf{0}_{m\times m} \end{bmatrix}$$

which is positive semi-definite.

- For any fixed $\lambda \in \mathbb{R}^{m(k-1)}, \mu \in \mathbb{R}^m$, the Lagrangian function is differentiable and convex over a non-empty open convex set X so that $(\hat{\mathbf{f}}, \hat{\eta})$ is an optimal solution to the minimization of $L(\mathbf{f}, \eta, \lambda, \mu)$ subject to $(\mathbf{f}, \eta) \in X$ if and only if $\nabla L(\hat{\mathbf{f}}, \hat{\eta}, \lambda, \mu) = \mathbf{0}$ if and only if

$$\hat{f}_y = \sum_{i=1, c(\omega_i)=y}^{m} \sum_{z=1, z\neq y}^{k} \lambda_{iz}\Phi(\omega_i) - \sum_{i=1, c(\omega_i)\neq y}^{m} \lambda_{iy}\Phi(\omega_i), y \in [1, k],$$

and

$$C = \mu_i + \sum_{z=1, z \neq c(\omega_i)}^{k} \lambda_{iz}, i \in [1, m].$$

– Note that for any fixed $\lambda \in \mathbb{R}^{m(k-1)}, \mu \in \mathbb{R}^m$, $C \neq \mu_i + \sum_{z=1, z \neq c(\omega_i)}^{k} \lambda_{iz}$ for some $i \in [1, m]$ if and only if the infimum of the Lagrangian function $L(\mathbf{f}, \eta, \lambda, \mu)$ is $-\infty$.

- Lagrangian dual function: for any $\lambda \in \mathbb{R}^{m(k-1)}, \mu \in \mathbb{R}^m$,

$$
\begin{aligned}
& \theta(\lambda, \mu) \\
= \; & \inf_{(\mathbf{f},\eta)\in X} L(\mathbf{f}, \eta, \lambda, \mu) \\
= \; & \begin{cases}
\frac{1}{2}\sum_{y=1}^{k} \|\hat{f}_y\|^2 + C\sum_{i=1}^{m}\hat{\eta}_i + \sum_{i=1}^{m}\sum_{y=1, y\neq c(\omega_i)}^{k} \\
\lambda_{iy}(1-\hat{\eta}_i - \langle \hat{f}_{c(\omega_i)} - \hat{f}_y, \Phi(\omega_i)\rangle_{\mathbb{H}_S})) - \sum_{i=1}^{m}\mu_i\hat{\eta}_i, \\
\qquad \text{if } C = \mu_i + \sum_{y=1,y\neq c(\omega_i)}^{k}\lambda_{iy}, i \in [1, m], \\
-\infty, \text{ otherwise}
\end{cases} \\
= \; & \begin{cases}
\sum_{i=1}^{m}\lambda_{ic(\omega_i)} - \frac{1}{2}\sum_{i,j=1}^{m}\sum_{y=1}^{k}(-1)^{\delta_{c(\omega_i)y}}\lambda_{iy} \\
\qquad (-1)^{\delta_{c(\omega_j)y}}\lambda_{jy}\langle\Phi(\omega_i), \Phi(\omega_j)\rangle_{\mathbb{H}}, \\
\text{if } C = \mu_i + \lambda_{ic(\omega_i)}, \lambda_{ic(\omega_i)} = \sum_{y=1,y\neq c(\omega_i)}^{k}\lambda_{iy}, i \in [1, m], \\
-\infty, \text{ otherwise.}
\end{cases}
\end{aligned}
$$

Lagrangian Dual Problem for Multi-Class SVM

$$
\begin{aligned}
\text{Maximize} \quad & \theta(\lambda, \mu) = \textstyle\sum_{i=1}^{m} \lambda_{ic(\omega_i)} - \frac{1}{2} \sum_{i,j=1}^{m} \sum_{y=1}^{k} \\
& (-1)^{\delta_{c(\omega_i)y}} \lambda_{iy} (-1)^{\delta_{c(\omega_j)y}} \lambda_{jy} K(\omega_i, \omega_j), \\
\text{Subject to} \quad & C = \mu_i + \lambda_{ic(\omega_i)},\ i \in [1, m], \\
& \lambda_{ic(\omega_i)} = \textstyle\sum_{y=1, y \neq c(\omega_i)}^{k} \lambda_{iy},\ i \in [1, m], \\
& \lambda_{iy} \geq 0,\ i \in [1, m],\ y \in [1, k], \\
& \mu_i \geq 0,\ i \in [1, m], \\
& (\lambda, \mu) \in \mathbb{R}^{mk} \times \mathbb{R}^{m}.
\end{aligned}
$$

Or equivalently,

$$
\begin{array}{ll}
\text{Maximize} & \theta(\lambda) = \sum_{i=1}^{m} \lambda_{ic(\omega_i)} - \frac{1}{2} \sum_{i,j=1}^{m} \sum_{y=1}^{k} \\
& (-1)^{\delta_{c(\omega_i)y}} \lambda_{iy} (-1)^{\delta_{c(\omega_j)y}} \lambda_{jy} K(\omega_i, \omega_j), \\
\text{Subject to} & \lambda_{iy} \geq 0,\ i \in [1, m],\ y \in [1, k], \\
& C - \lambda_{ic(\omega_i)} \geq 0,\ i \in [1, m], \\
& \lambda_{ic(\omega_i)} - \sum_{y=1, y \neq c(\omega_i)}^{k} \lambda_{iy} = 0,\ i \in [1, m], \\
& \lambda \in \mathbb{R}^{mk}.
\end{array}
$$

- A quadratic programming (QP) problem.

Qualification of the Dual Problem

- The object function

$$\theta(\lambda) = \sum_{i=1}^{m} \lambda_{ic(\omega_i)} - \frac{1}{2} \sum_{i,j=1}^{m} \sum_{y=1}^{k} (-1)^{\delta_{c(\omega_i)y}} \lambda_{iy} (-1)^{\delta_{c(\omega_j)y}} \lambda_{jy} K(\omega_i, \omega_j)$$

  is infinitely differentiable and concave so that it is pseudoconcave at any feasible point.

- The inequality constraint functions $g_{iy}(\lambda) = \lambda_{iy}$, $i \in [1, m]$, $y \in [1, k]$, $\tilde{g}_i(\lambda) = C - \lambda_{ic(\omega_i)}, i \in [1, m]$, and the equality constraint function $h_i(\lambda) = \sum_{y=1}^{k} (-1)^{\delta_{c(\omega_i)y}} \lambda_{iy}$, $i \in [1, m]$, are affine functions so that they are infinitely differentiable, concave and convex and then quasiconcave and quasiconvex at any feasible point.

- Any feasible point λ which satisfies the Kuhn-Tucker necessary conditions is a global maximum solution.

Justification of Strong Duality for Multi-Class Kernel-Based SVM

- $X = \mathbb{H}_S^k \times \mathbb{R}^m$: a non-empty convex set.
- $F(\mathbf{f}, \eta) = \frac{1}{2} \sum_{y=1}^{k} \|f_y\|^2 + C \sum_{i=1}^{m} \eta_i$: a convex function on X.
- $g_{iy}(\mathbf{f}, \eta) = 1 - \eta_i - \langle f_{c(\omega_i)} - f_y, \Phi(\omega_i) \rangle_{\mathbb{H}_S}, i \in [1, m], y \in [1, k], y \neq c(\omega_i)$: affine functions so that they are convex functions on X.
- $h_i(\mathbf{f}, \eta) = -\eta_i, 1 \leq i \leq m$: affine functions so that they are convex functions on X.
- There exists an $(\mathbf{f}', \eta') \in X$ such that $\mathbf{g}(\mathbf{f}', \eta') < \mathbf{0}$ and $\mathbf{h}(\mathbf{f}', \eta') < \mathbf{0}$.

Then we have

$$\begin{aligned} & \inf\{F(\mathbf{f},\eta) : (\mathbf{f},\eta) \in X, \mathbf{g}(\mathbf{f},\eta) \le \mathbf{0}, \mathbf{h}(\mathbf{f},\eta) \le \mathbf{0}\} \\ = \quad & \sup\{\theta(\lambda,\mu) : (\lambda,\mu) \ge \mathbf{0}\}. \end{aligned}$$

- For a non-trivial labeled training sample, the inf is finite and can be achieved at some feasible point $(\mathbf{f}^{SVM}, \eta^{SVM})$. Then $\sup\{\theta(\lambda,\mu) \mid (\lambda,\mu) \ge \mathbf{0}\}$ is achieved at some $(\lambda^{SVM}, \mu^{SVM}) \ge \mathbf{0}$.
- The primal and dual problems are equivalent.

The Multi-Class Kernel-Based SVM Algorithm

- $S = (\omega_1, \ldots, \omega_m)$: a non-trivial labeled training sample of size m with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- $h_S^{SVM} : \mathscr{I} \to \mathscr{Y}$: the hypothesis returned by the multi-class kernel-based SVM such that for all $\omega \in \mathscr{I}$,

$$\begin{aligned} h_S^{SVM}(\omega) &= \arg\max_{y \in \mathscr{Y}} \left\langle f_y^{SVM}, \Phi(\omega) \right\rangle_{\mathbb{H}} \\ &= \arg\max_{y \in \mathscr{Y}} \left\langle \sum_{i=1}^{m} (-1)^{\delta_{c(\omega_i)y}} \lambda_{iy}^{SVM} \Phi(\omega_i), \Phi(\omega) \right\rangle_{\mathbb{H}} \\ &= \arg\max_{y \in \mathscr{Y}} \sum_{i=1}^{m} (-1)^{\delta_{c(\omega_i)y}} \lambda_{iy}^{SVM} K(\omega_i, \omega) \end{aligned}$$

where the returned weight vectors are

$$f_y^{SVM} = \sum_{i=1}^{m} (-1)^{\delta_{c(\omega_i)y}} \lambda_{iy}^{SVM} \Phi(\omega_i).$$

- The hypothesis solution h_S^{SVM} depends only on kernel values between items and not directly on the feature vectors of items.

AdaBoost.MH : Multi-Class Hamming Loss

- AdaBoost.MH is a boosting algorithm for multi-class classification.
- AdaBoost.MH applies to the multi-label setting where the label space $\mathscr{Y}$ is $\{-1, +1\}^k$.
- As in the binary case, it returns a convex combination of base classifiers selected from a hypothesis set $\mathcal{H}$.
- AdaBoost.MH reduces a multi-label training data of size m to a binary training data of size mk by splitting the ith multi-labeled item to k binary-labeled items as follows:

$$(\omega_i, c(\omega_i)) \to ((\omega_i, 1), c(\omega_i)_1), \ldots, ((\omega_i, k), c(\omega_i)_k), \;\; i \in [1, m].$$

- The base classifiers are functions mapping from $\mathscr{I} \times \{1, 2, \ldots, k\}$ to $\{-1, +1\}$.

- Adaboost.MH maintains a distribution on the double index set $[1, m] \times [1, k]$ which will be updated at each round of boosting. The initial distribution D_1 is set to be the uniform distribution, i.e., $D_1(i, l) = 1/(mk)$ for all $i \in [1, m], l \in [1, k]$. Let D_t be the distribution on $[1, m] \times [1, k]$ at the t-th round of boosting.

- At the t-th round of boosting, the base classifier $h_{S,t}$ is selected that minimizes the error on the training sample weighted by the distribution D_t:

$$\begin{aligned} h_{S,t} &\in \arg\min_{h \in \mathcal{H}} \underset{(i,l)\sim D_t}{P} (h(\omega_i, l) \neq c(\omega_i)_l) \\ &= \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{m} \sum_{l=1}^{k} D_t(i, l) 1_{h(\omega_i, l) \neq c(\omega_i)_l}. \end{aligned}$$

  - Instead of a hypothesis with minimal weighted error, $h_{S,t}$ can be more generally the base classifier returned by a weak learning algorithm trained on the distribution D_t.

- Thus, AdaBoost.MH working on a multi-labeled sample $S = ((\omega_1, c(\omega_1)), \ldots, (\omega_m, c(\omega_m)))$ of size m is equivalent to AdaBoost working on a binary-labeled sample $S' = (((\omega_1, 1), c(\omega_1)_1), \ldots, ((\omega_m, k), c(\omega_m)_k))$ of size mk.

- The complexity of the AdaBoost.MH algorithm is that of the AdaBoost applied to a sample of size mk. For $\mathscr{I} \subseteq \mathbb{R}^N$, using boosting stumps as base classifiers, the complexity of the algorithm is therefore in $O((mk)\ln(mk) + mkNT)$. Thus, for a large number k of labels, the algorithm may become impractical using a single processor.

- The weak learning condition for the application of AdaBoost in this scenario requires that at each round there exists a base classifier $h_{S,t} : \mathscr{I} \times \{1, 2, \ldots, k\} \to \{-1, +1\}$ such that $\underset{(i,l)\sim D_t}{P}(h_{S,t}(\omega_i, l) \neq c(\omega_i)_l) \leq \frac{1}{2} - \gamma$. This may be hard to achieve if classes are close and it is difficult to distinguish them.

The AdaBoost.MH Algorithm for $\mathcal{H} \subseteq \left(\{-1,+1\}^k\right)^{\mathscr{I}}$

$\textsc{AdaBoost.MH}\ (S = ((\omega_1, c(\omega_1)), \ldots, (\omega_m, c(\omega_m))))$

```
for i ← 1 to m do
    for l ← 1 to k do
        D_1(i, l) ← 1/(mk)
for t ← 1 to T do
    h_{S,t} ← base classifier in H with small error
```

$$\epsilon_t = \underset{(i,l)\sim D_t}{P}(h_{S,t}(\omega_i, l) \neq c(\omega_i)_l) = \sum_{i=1}^{m}\sum_{l=1}^{k} D_t(i,l) 1_{h_{S,t}(\omega_i,l)\neq c(\omega_i)_l}$$

```
    α_t ← (1/2) log((1 − ε_t)/ε_t)
    Z_t ← 2[ε_t(1 − ε_t)]^{1/2}        ▷ normalization factor
```

```
        for i ← 1 to m do
            for l ← 1 to k do
            D_{t+1}(i,l) ← D_t(i,l) exp(−α_t h_{S,t}(ω_i,l) c(ω_i)_l) / Z_t
g_S ← Σ_{t=1}^{T} α_t h_{S,t}
return h_S = sgn(g_S)
```

$$\text{for } i \leftarrow 1 \textbf{ to } m \textbf{ do}$$

$$\quad \textbf{for } l \leftarrow 1 \textbf{ to } k \textbf{ do}$$

$$\quad\quad D_{t+1}(i,l) \leftarrow \frac{D_t(i,l)\exp(-\alpha_t h_{S,t}(\omega_i,l)c(\omega_i)_l)}{Z_t}$$

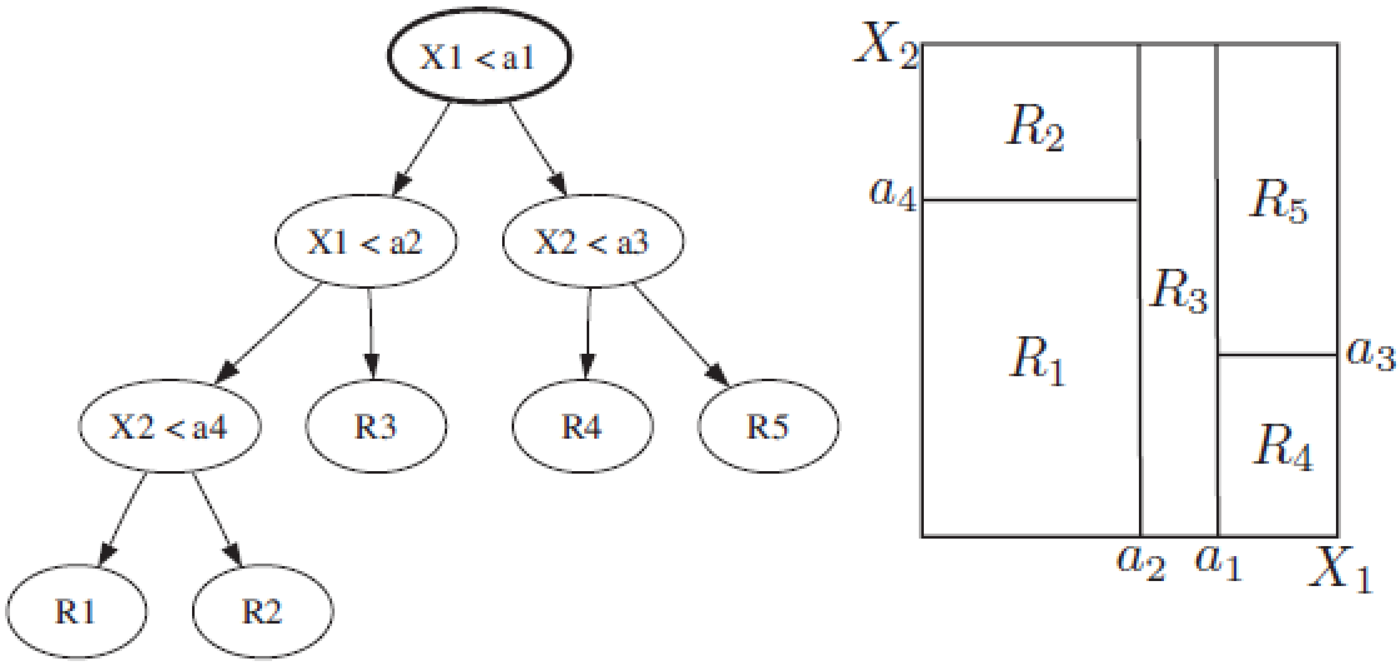
$$g_S \leftarrow \sum_{t=1}^{T} \alpha_t h_{S,t}$$

$$\textbf{return } h_S = \mathrm{sgn}(g_S)$$

Definition 8.1: Binary Decision Trees

- $\mathscr{I}$: the input space associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$.
- $\mathscr{Y}$: the label space.
- $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$: an N-dimensional feature space.
  - Each feature set $\mathcal{X}_i$ is either a numerical set or a categorical set.
- $\mathbf{X} = (X_1, X_2, \ldots, X_N) : \mathscr{I} \to \mathcal{X}$: a measurable feature mapping which associates an item ω with a feature vector $\mathbf{X}(\omega)$, where $X_i(\omega)$ is called the ith feature of item ω.
  - With the probability space $(\mathscr{I}, \mathcal{F}, P)$, $\mathbf{X}$ is a random vector and each feature variable X_i is a random variable.

A binary decision tree is a tree representation of a partition of the feature space $\mathcal{X}$. Each interior node of a decision tree corresponds to a question related to features. It can be a numerical question of the form $X_i \leq a$ for a continuous feature variable X_i and some threshold $a \in \mathbb{R}$ or a categorical question such as $X_i \in \{\text{blue, white, red}\}$, when feature X_i takes a categorical value such as a color. Each leaf is labeled with a label $l \in \mathscr{Y}$.



Left: An example of a binary decision tree with numerical quesions based on two variables X_1 and X_2.
Right: The partition of the two-dimensional feature space induced by that decision tree.

Remarks

- Binary decision trees can be defined using more complex node questions, resulting in partitions based on more complex decision surfaces.
  - Binary space partition (BSP) trees partition the space with convex polyhedral regions, based on questions of the form $\sum_{j=1}^{N} \alpha_j X_j \leq b$.
  - Binary sphere trees partition with pieces of spheres based on questions of the form $\|\mathbf{X} - \mathbf{a}_0\| \leq r$, where $\mathbf{X}$ is a feature vector, $\mathbf{a}_0$ a fixed vector, and r a fixed positive real number.
- More complex tree questions lead to richer partitions and thus hypothesis sets, which can cause overfitting in the absence of a sufficiently large training sample. They also increase the computational complexity of prediction and training.

- Binary decision trees can also be generalized to branching factors greater than two, but binary trees are most commonly used due to computational considerations.
- Each leaf defines a region of the feature space $\mathcal{X}$ formed by the set of items corresponding exactly to the same node responses and thus the same traversal of the tree.
  - By definition, no two regions intersect and each item belongs to exactly one region.
  - Thus, leaf regions define a partition of the feature space $\mathcal{X}$.
- In multi-class classification, the label of a leaf is determined using the training sample: the class with the majority representation among the training items falling in a leaf region defines the label of that leaf, with ties broken arbitrarily.

Label Prediction by a Binary Decision Tree

- To predict the label of any item $\omega \in \mathscr{I}$, we take the feature vector $\mathbf{X}(\omega)$ and start at the root node of the binary decision tree and go down the tree until a leaf is found, by moving to the right child of a node when the response to the node question is positive, and to the left child otherwise. When we reach a leaf, we associate ω with the label of this leaf.

Training a Binary Decision Tree with a Greedy Algorithm

- The greedy algorithm consists of starting with a tree reduced to a single (root) node, which is a leaf whose label is the class that has majority over the entire sample.
- Next, at each round, a node n is split based on some question q. The pair (n, q) is chosen so that the node impurity is maximally decreased according to some measure of impurity $F(n)$ of a node n.
- The decrease in node impurity after a split of node n based on question q is defined as follows:
  - $n_{+}(n, q)$: the right child of n after the split.
  - $n_{-}(q, n)$: the left child of n after the split.
  - $\eta(n, q)$: the fraction of the items in the region defined by n that are moved to $n_{-}(n, q)$.

The total impurity of the leaves $n_-(n,q)$ and $n_+(n,q)$ is therefore

$$\eta(n,q)F(n_-(n,q)) + (1-\eta(n,q))F(n_+(n,q)).$$

Thus, the decrease in impurity $\tilde{F}(n,q)$ by that split is given by

$$\tilde{F}(n,q) = F(n) - (\eta(n,q)F(n_-(n,q)) + (1-\eta(n,q))F(n_+(n,q))).$$

- In practice, the algorithm is stopped once all leaves have reached a sufficient level of purity, when the number of items per leaf has become too small for further splitting or based on some other similar heuristic.
- The general problem of determining partition with minimum empirical error is NP-hard.

Greedy Algorithm for Building a Binary Decision Tree

GreedyDecisionTree($S = ((\omega_1, c(\omega_1)), \ldots, (\omega_m, c(\omega_m))))$

tree $\leftarrow \{n_0\}$ $\quad \triangleright$ root node

for $t \leftarrow 1$ **to** T **do**

$\qquad (n_t, q_t) \leftarrow \arg\max_{(n,q)} \tilde{F}(n, q)$

$\qquad$ Split(tree, n_t, q_t)

return tree

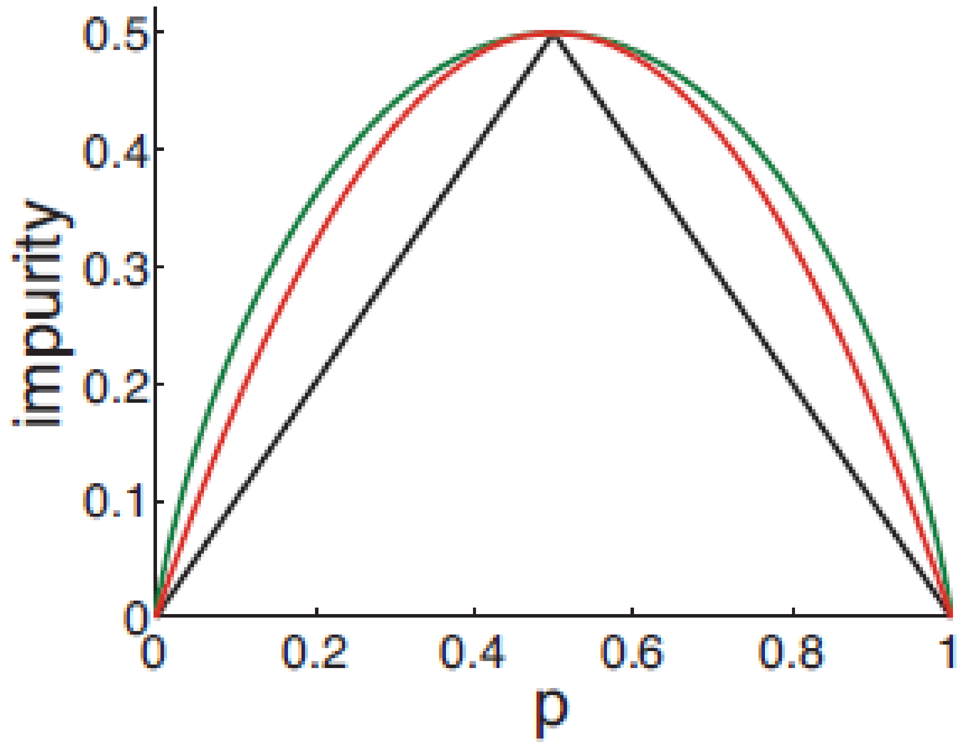
Three Most Commonly Used Measures of Node Impurity

Let

- $p_l(n)$: the fraction of items at a node n that belong to class $l \in [1, k]$.

The three commonly used measures of impurity $F(n)$ of a node n are

- Misclassification : $F(n) = 1 - \max_{l \in [1,k]} p_l[n]$.
- Entropy : $F(n) = -\sum_{l=1}^{k} p_l[n] \log_2 p_l[n]$.
- Gini index : $F(n) = \sum_{l=1}^{k} p_l[n](1 - p_l[n])$.



Three most commonly used measures of node impurity as functions of the fraction of positive examples in the binary case: misclassification (in black), entropy (in green, scaled by 0.5), and the Gini index (in red).

Remarks

- All three impurity functions depend only on the class distribution $\{p_l(n), i \in [1, k]\}$ of items at a node n.
- The entropy and Gini index impurity functions are upper bounds on the misclassification impurity function.
- All three functions are concave, which ensures that

$$F(n) - (\eta(n,q)F(n_-(n,q)) + (1-\eta(n,q))F(n_+(n,q))) \geq 0.$$

- However, the misclassification function is piecewise linear, so $\tilde{F}(n,q)$ is zero if the fraction of positive items, when $k = 2$ so that each item is labeled with -1 or $+1$, remains less than (or more than) half after a split. In some cases, the impurity cannot be decreased by any split using that criterion.
- In contrast, the entropy and Gini functions are strictly convex,

which guarantees a strict decrease in impurity. Furthermore, they are differentiable which is a useful feature for numerical optimization. Thus, the Gini index and the entropy criteria are typically preferred in practice.

Node Questions for Binary Decision Trees

- $S = ((\omega_1, c(\omega_1)), \ldots, (\omega_m, c(\omega_m)))$: a labeled sample of size m.
- $S_n \triangleq \{\omega_i, i \in [1, m] \mid \omega_i \text{ is in the region defined by the node } n\}$: the reduced sample to a node n.
- $X_j \leq a$: numerical questions for a continuous feature variable $X_j \in \mathcal{X}_j$.
  - The threshold value a are selected from the set $\{X_j(\omega_i), i \in [1, m] \mid \omega_i \in S_n\}$.
- $X_j \in A$: categorical questions for a categorical feature variable $X_j \in \mathcal{X}_j$.
  - The set A is any subset of $\mathcal{X}_j$ with size no more than half of $|\mathcal{X}_j|$.

Issues of the Greedy Algorithm

- The greedy nature of the algorithm: a seemingly bad split may dominate subsequent useful splits, which could lead to trees with less impurity overall.
  - This can be overcome to a certain extent by using a look-ahead of some depth d to determine the splitting decisions, but such look-aheads can be computationally very costly.
- To achieve some desired level of impurity, trees of relatively large sizes may be needed. But larger trees define overly complex hypotheses with high VC-dimensions (see Exercise 8.5) and thus could overfit.

Training a Binary Decision Tree with a Grow-Then-Prune Strategy

- First a very large tree is grown until it fully fits the training sample or until no more than a very small number of items are left at each leaf.
- Then, the resulting tree, denoted as *tree*, is pruned back to minimize an objective function,

$$G_\lambda(tree) = \sum_{n \in L_{tree}} |n| F(n) + \lambda |L_{tree}|,$$

  defined based on generalization bounds as the sum of an empirical error and a complexity term that can be expressed in terms of the size of L_{tree}, the set of leaves of the *tree*.
  - $|n|$: the size of the region defined by the node n.

  - $\lambda > 0$: a regularization parameter determining the trade-off between misclassification, or more generally impurity, versus tree complexity.
  - λ is determined by n-fold cross-validation.
- $\hat{R}(\mathit{tree}') = \sum_{n \in L_{\mathit{tree}'}} |n| F(n)$: the total empirical error of a tree tree'.
- We seek a sub-tree tree_λ of the tree that minimizes G_λ and that has the smallest size.
  - tree_λ can be shown to be unique.
- To determine tree_λ, the following pruning method is used, which defines a finite sequence of nested sub-trees $\mathit{tree}^{(0)}, \ldots,$ $\mathit{tree}^{(n)}$.
- We start with the full tree $\mathit{tree}^{(0)} = \mathit{tree}$ and for any $i \in [0, n-1]$, define $\mathit{tree}^{(i+1)}$ from $\mathit{tree}^{(i)}$ by collapsing an internal node n' of $\mathit{tree}^{(i)}$, that is by replacing the sub-tree tree'

rooted at n' with a leaf, or equivalently by combining the regions of all the leaves dominated by n'.

– n' is chosen so that collapsing it causes the smallest per node increase in $\hat{R}(tree^{(i)})$, that is the smallest $r(tree^{(i)}, n')$ defined by

$$r(tree^{(i)}, n') = \frac{|n'|F(n') - \hat{R}(tree')}{L_{tree'} - 1}.$$

– If several nodes n' in $tree^{(i)}$ cause the same smallest increase per node $r(tree^{(i)}, n')$, then all of them are pruned to define $tree^{(i+1)}$ from $tree^{(i)}$.

– This procedure continues until the tree $tree^{(n)}$ obtained has a single node.

– The optimal sub-tree tree_λ can be shown to be among the elements of the sequence $tree^{(0)}$, ..., $tree^{(n)}$.