EE6550 Machine Learning

Lecture Eight – Regression

Chung-Chin Lu
Department of Electrical Engineering
National Tsing Hua University

May 1, 2017

Regression Problem

- $\mathscr{I}$: the input space of all items, associated with a probability space $(\mathscr{I}, \mathcal{F}, D)$.
- $\mathscr{Y}' = \mathscr{Y} \subseteq \mathbb{R}$: the output and label spaces, which typically are $[-M, M]$ for some $M > 0$ or $\mathbb{R}$.
- $c : \mathscr{I} \to \mathscr{Y}$: a fixed but unknown target concept in the concept class $\mathcal{C}$.
  - c is a real-valued measurable function on $\mathscr{I}$.
- $\mathcal{H}$: a hypothesis set of real-valued measurable functions on $\mathscr{I}$.

- $L : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}^+ = [0, \infty)$: a measurable loss function, typically chosen as the squared-error loss function $L(y', y) = (y' - y)^2$.
  - The loss function can be chosen as $L(y', y) = |y' - y|^p$ for some $p \geq 1$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the input space according to D, with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- Problem: find a hypothesis $h : \mathscr{I} \to \mathscr{Y}$ in $\mathcal{H}$ with small generalization error w.r.t. the target concept c,

$$R(h) = \mathop{E}_{\omega \sim D}[L(h(\omega), c(\omega))].$$

  - When $L(y', y) = (y' - y)^2$ is the squared-error loss function, $R(h) = \mathop{E}_{\omega \sim D}[(h(\omega) - c(\omega))^2]$ is called the mean squared-error (MSE) of the hypothesis h w.r.t. the target concept c.

The Contents of This Lecture

- Generalization bounds
- Linear regression
- Kernel ridge regression
- Support vector regression
- Lasso

Uniformly Bounded Regression Problem

- Assumption : in this lecture we will assume that

$$L(h(\omega), c(\omega)) \leq M$$

for some $M > 0$ for all $\omega \in \mathscr{I}$ and for all $h \in \mathcal{H}$ and for all $c \in C$.

Regression Generalization Bound - Finite Hypothesis Set

Theorem 10.1: Let

- C : a concept class C to learn
- $\mathcal{H}$: a finite hypothesis set
- L: a measurable loss function such that $L(h(\omega), c(\omega)) \leq M$ for all $\omega \in \mathscr{I}$ and for all h in $\mathcal{H}$, for all c in C.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the input space $\mathscr{I}$ according to a distribution D, with labels $(c(\omega_1), \ldots, c(\omega_m))$ by a fixed but unknown concept c.

For any $\delta > 0$, with probability at leat $1 - \delta$, the following generalization error bound holds for all $h \in \mathcal{H}$ for any target $c \in C$ and for any distribution D:

$$R(h) \leq \hat{R}_S(h) + M\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{2m}}$$

and

$$|R(h) - \hat{R}_S(h)| \leq M\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2m}}.$$

Proof. For any $h \in \mathcal{H}$ and any $c \in C$, we have

- $\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} L(h(\omega_i), c(\omega_i))$: the empirical error which is a measurable function of the random sample S.

- $\underset{S\sim D^m}{E}[\hat{R}_S(h)] = \underset{S\sim D^m}{E}[\frac{1}{m}\sum_{i=1}^{m} L(h(\omega_i), c(\omega_i))] = \frac{1}{m}\sum_{i=1}^{m} \underset{S\sim D^m}{E}[L(h(\omega_i), c(\omega_i))] = \underset{\omega\sim D}{E}[L(h(\omega), c(\omega))] = R(h)$.

- $L(h(\omega_1), c(\omega_1)), \ldots, L(h(\omega_m), c(\omega_m))$: independent r.v.'s taking values in $[0, M]$

By Hoeffding's inequality, we have

$$\underset{S\sim D^m}{P}(R(h) - \hat{R}_S(h) > \epsilon) \quad \leq \quad e^{2m\epsilon^2/M^2}, \tag{1}$$

$$\underset{S\sim D^m}{P}(R(h) - \hat{R}_S(h) < -\epsilon) \quad \leq \quad e^{2m\epsilon^2/M^2} \tag{2}$$

for any $\epsilon > 0$. Now we have

$$\begin{aligned}
& \underset{S\sim D^m}{P}(\max_{h\in\mathcal{H}} R(h) - \hat{R}_S(h) > \epsilon) \\
= \quad & \underset{S\sim D^m}{P}(\cup_{h\in\mathcal{H}}(R(h) - \hat{R}_S(h) > \epsilon)) \\
\leq \quad & \sum_{h\in\mathcal{H}} \underset{S\sim D^m}{P}(R(h) - \hat{R}_S(h) > \epsilon) \text{ by union bound} \\
\leq \quad & |\mathcal{H}| e^{-2m\epsilon^2/M^2} \text{ by Eq. (1).}
\end{aligned}$$

Setting $\delta = |\mathcal{H}| e^{-2m\epsilon^2/M^2}$ and solving $\epsilon = M\sqrt{\frac{\ln|\mathcal{H}|+\ln\frac{1}{\delta}}{2m}}$, we have

$$\underset{S\sim D^m}{P}\left(\max_{h\in\mathcal{H}} R(h) - \hat{R}_S(h) > M\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{2m}}\right) < \delta.$$

Thus with probability at least $1-\delta$,

$$\forall\ h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + M\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{2m}},$$

for any target $c \in C$ and for any distribution D. Furthermore, if we let $\frac{\delta}{2} = |\mathcal{H}|e^{-2m\epsilon^2/M^2}$ and solve $\epsilon = M\sqrt{\frac{\ln|\mathcal{H}|+\ln\frac{2}{\delta}}{2m}}$, we have

$$\underset{S\sim D^m}{P}\left(\max_{h\in\mathcal{H}} R(h) - \hat{R}_S(h) > M\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2m}}\right) < \frac{\delta}{2}$$

and

$$\underset{S\sim D^m}{P}\left(\max_{h\in\mathcal{H}} R(h) - \hat{R}_S(h) < -M\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2m}}\right) < \frac{\delta}{2}$$

Thus with probability at least $1-\delta$,

$$\forall\ h \in \mathcal{H},\ \ |R(h) - \hat{R}_S(h)| \leq M\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2m}},$$

for any target $c \in C$ and for any distribution D. $\square$

Remarks

- The regression generalization bound $\hat{R}_S(h) + M\sqrt{\frac{\ln|\mathcal{H}|+\ln\frac{1}{\delta}}{2m}}$ suggests seeking a trade-off between reducing the empirical error versus controlling the size of the hypothesis set.
  - A larger hypothesis set is penalized by the second term but could help reduce the empirical error, that is the first term.
  - Occam's Razor principle (law of parsimony): the simplest explanation is best. Thus if all other things being equal (a similar empirical error), a simpler (smaller) hypothesis set is better.
- The regression generalization bound is in $O(\sqrt{\frac{\ln|\mathcal{H}|}{m}})$, not in $O(\frac{\ln|\mathcal{H}|}{m})$.

$\phi_p(x) = |x|^p$ **Is a** (pM^{p-1})**-Lipschitz Function on** $[-M, M]$

Consider the ratio $\frac{|\phi_p(x)-\phi_p(y)|}{|x-y|}$, $x, y \in [-M, M]$ and $x \neq y$. If $|x| = |y|$, then the ratio is zero since ϕ_p is an even function. Consider $|x| \neq |y|$ and without loss of generality, assume $|x| > |y|$. If $xy < 0$, then $|x - y| > |x| - |y|$ and

$$\frac{|\phi_p(x) - \phi_p(y)|}{|x - y|} < \frac{|\phi_p(|x|) - \phi_p(|y|)|}{|x| - |y|} = \frac{\phi_p(|x|) - \phi_p(|y|)}{|x| - |y|}$$

since $\phi_p(x) = x^p$ is an increasing function on $\mathbb{R}^+$. If $xy \geq 0$, then $|x - y| = |x| - |y|$ and

$$\frac{|\phi_p(x) - \phi_p(y)|}{|x - y|} = \frac{|\phi_p(|x|) - \phi_p(|y|)|}{|x| - |y|} = \frac{\phi_p(|x|) - \phi_p(|y|)}{|x| - |y|}.$$

Thus it is sufficient to consider $\phi_p(x) = x^p$ on $[0, M]$. By the mean-value theorem, we have for $0 \le y < x \le M$,

$$\frac{\phi_p(x) - \phi_p(y)}{x - y} = \phi_p'(z) = pz^{p-1} \le pM^{p-1}$$

for some $z \in (y, x)$. We conclude that

$$\frac{|\phi_p(x) - \phi_p(y)|}{|x - y|} \le pM^{p-1}, \ \forall \ x, y \in [-M, M], \ x \neq y$$

and then

$$|\phi_p(x) - \phi_p(y)| \le pM^{p-1}|x - y|, \ \forall \ x, y \in [-M, M].$$

Thus $\phi_p(x) = |x|^p$ is a (pM^{p-1})-Lipschitz function on $[-M, M]$. $\square$

Rademacher Complexity of L_p Loss Functions

Theorem 10.2: Let

- c : a fixed but unknown concept in a concept class C of real-valued measurable functions on the input space $\mathscr{I}$ to learn.
- $\mathcal{H}$: a hypothesis set of real-valued measurable functions on $\mathscr{I}$, which may be an infinite set, such that

$$|h(\omega) - c(\omega)| \le M, \ \forall\, \omega \in \mathscr{I}, \ \forall\, h \in \mathcal{H}.$$

- $L_p(y', y) = |y' - y|^p$: the L_p loss function, $p \ge 1$.
- $\mathcal{H}_p = \{L_p(h, c) \mid h \in \mathcal{H}\}$: the family of all real-valued measurable functions $L_p(h(\omega), c(\omega))$ on $\omega \in \mathscr{I}$ with $h \in \mathcal{H}$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the input space $\mathscr{I}$ according to a distribution D, with labels $(c(\omega_1), \ldots, c(\omega_m))$ by a fixed but unknown concept c.

Then we have

$$\hat{\mathfrak{R}}_S(\mathcal{H}_p) \le pM^{p-1}\hat{\mathfrak{R}}_S(\mathcal{H}).$$

Proof.

- $\phi_p : [-M, M] \to \mathbb{R}$ with $\phi_p(x) = |x|^p$: a (pM^{p-1})-Lipschitz function on $[-M, M]$.
- $\mathcal{H}' = \{h - c \mid h \in \mathcal{H}\}$: $|h'(\omega)| \le M$ for all $\omega \in \mathscr{I}$ and for all $h' \in \mathcal{H}'$.
- $\mathcal{H}_p = \phi_p \circ \mathcal{H}'$.
- $\sup_{h' \in \mathcal{H}'} \left(\sum_{i=1}^{j} \sigma_i (\phi_p \circ h')(\omega_i) + \sum_{i=j+1}^{m} (pM^{p-1}) \sigma_i h'(\omega_i) \right)$ is finite for all $\sigma_i \in \{-1, +1\}, i \in [1, m]$ and for all $j \in [0, m]$.

By Talagrand's lemma (Lemma 4.2), we have

$$\hat{\mathfrak{R}}_S(\mathcal{H}_p) = \hat{\mathfrak{R}}_S(\phi_p \circ \mathcal{H}') \le pM^{p-1}\hat{\mathfrak{R}}_S(\mathcal{H}').$$

But

$$
\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{H}') &= \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h' \in \mathcal{H}'} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h'(\omega_i) \\
&= \mathop{E}_{\sigma} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (h(\omega_i) - c(\omega_i)) \right] \\
&= \mathop{E}_{\sigma} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(\omega_i) \right] + \mathop{E}_{\sigma} \left[\frac{1}{m} \sum_{i=1}^{m} (-\sigma_i) c(\omega_i) \right] \\
&= \hat{\mathfrak{R}}_S(\mathcal{H}) + \frac{1}{m} \sum_{i=1}^{m} \mathop{E}_{\sigma} [\sigma_i]\, c(\omega_i) \\
&= \hat{\mathfrak{R}}_S(\mathcal{H})
\end{aligned}
$$

since $\mathop{E}_{\sigma} [\sigma_i] = 0$ for all $i \in [1, m]$. $\square$

Rademacher Complexity Regression Generalization Bounds

Theorem 10.3: Let

- c : a fixed but unknown concept in a concept class C of real-valued measurable functions on the input space $\mathscr{I}$ to learn.
- $\mathcal{H}$: a hypothesis set of real-valued measurable functions on $\mathscr{I}$, which may be an infinite set, such that

$$|h(\omega) - c(\omega)| \leq M, \ \forall\, \omega \in \mathscr{I}, \ \forall\, h \in \mathcal{H}.$$

- $L_p(y', y) = |y - y|^p$: the L_p loss function, $p \geq 1$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the input space $\mathscr{I}$ according to a distribution D with labels $(c(\omega_1), \ldots, c(\omega_m))$.

For any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all h in $\mathcal{H}$:

$$
\begin{aligned}
\underset{\omega\sim D}{E}[|h(\omega)-c(\omega)|^p] &\leq \frac{1}{m}\sum_{i=1}^{m}|h(\omega_i)-c(\omega_i)|^p + 2pM^{p-1}\mathfrak{R}_m(\mathcal{H}) \\
&\quad + M^p\sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \\
\underset{\omega\sim D}{E}[|h(\omega)-c(\omega)|^p] &\leq \frac{1}{m}\sum_{i=1}^{m}|h(\omega_i)-c(\omega_i)|^p + 2pM^{p-1}\hat{\mathfrak{R}}_S(\mathcal{H}) \\
&\quad + 3M^p\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}.
\end{aligned}
$$

Proof. Apply the Rademacher complexity bound (Theorem 3.1) to the following family:

- $\mathcal{H}_p = \{|h-c|^p \mid h \in \mathcal{H}\}$: the family of all real-valued measurable functions $|h(\omega)-c(\omega)|^p$ from $\omega \in \mathscr{I}$ to $[0, M^p]$

with $h \in \mathcal{H}$.

Now we have

$$
\begin{aligned}
\underset{\omega \sim D}{E}[|h(\omega) - c(\omega)|^p] &\leq \frac{1}{m}\sum_{i=1}^{m}|h(\omega_i) - c(\omega_i)|^p + 2\mathfrak{R}_m(\mathcal{H}_p) \\
&\quad + M^p\sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \\
\underset{\omega \sim D}{E}[|h(\omega) - c(\omega)|^p] &\leq \frac{1}{m}\sum_{i=1}^{m}|h(\omega_i) - c(\omega_i)|^p + 2\hat{\mathfrak{R}}_S(\mathcal{H}_p) \\
&\quad + 3M^p\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}.
\end{aligned}
$$

The proof is complete by applying Theorem 10.2. $\square$

Remarks

- These Rademacher complexity regression generalization bounds suggest a trade-off between reducing the empirical error, which may require more complex hypothesis sets, and controlling the Rademacher complexity of $\mathcal{H}$, which may increase the empirical error.
- An important benefit of the second learning bound is that it is data dependent. This can lead to more accurate learning guarantees.

- If $\mathcal{H} = \{\omega \mapsto \langle f, \Phi(\omega)\rangle_{\mathbb{H}} + b \mid f \in \mathbb{H} \text{ with } \|f\|_{\mathbb{H}} \leq \Lambda, b \leq r\Lambda\}$ is a kernel-based hypothesis set, where $\mathbb{H}$ and Φ are the RKHS and the associated feature mapping of a PDS kernel K over the input space $\mathscr{I}$ with $K(\omega, \omega) \leq r^2 \ \forall\, \omega \in \mathscr{I}$, its empirical Rademacher complexity w.r.t. the sample S can be bounded by

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{\Lambda\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + \frac{r\Lambda}{\sqrt{m}} \leq 2\sqrt{\frac{r^2\Lambda^2}{m}}$$

  from Theorem 5.5 of Lecture 4, where $\mathbf{K} = [K(\omega_i, \omega_j)]$ is the $m \times m$ kernel matrix associated to the kernel K and the sample S.

- As discussed for binary classification:
  - Estimating the Rademacher complexity may be computationally hard for some $\mathcal{H}$'s.
  - Is there a combinatorial measure that is easier to compute?

Definition 10.1: Shattering

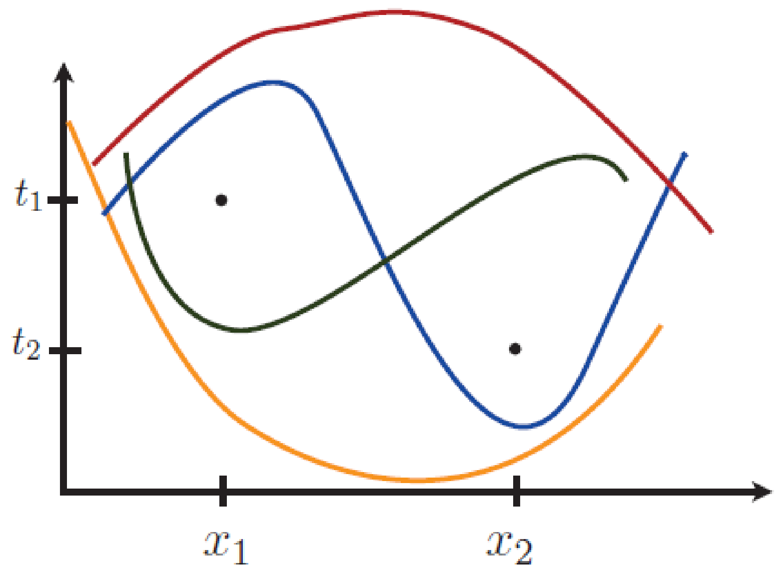
- $\mathscr{I}$: the input space of all possible items.
- $\mathcal{G}$: a family of measurable functions from $\mathscr{I}$ to $\mathbb{R}$.
- $S = \{\omega_1, \ldots, \omega_m\}$: an m-subset of $\mathscr{I}$.

S is said to be shattered by $\mathcal{G}$ if there are $t_1, \ldots, t_m \in \mathbb{R}$ such that

$$\left| \left\{ \begin{bmatrix} \operatorname{sgn}(g(\omega_1) - t_1) \\ \vdots \\ \operatorname{sgn}(g(\omega_m) - t_m) \end{bmatrix} \;\middle|\; g \in \mathcal{G} \right\} \right| = 2^m.$$

- When exist, the thresholds $t_1, \ldots, t_m$ are said to witness the shattering.

- Thus, $S = \{\omega_1, \ldots, \omega_m\}$ is shattered if for some witnesses $t_1, \ldots, t_m \in \mathbb{R}$, the family $\mathcal{G}$ is rich enough to contain a measurable function whose graph in the $\mathscr{I} \times \mathbb{R}$ plane goes above a subset A of the set of points $I = \{(\omega_i, t_i) \mid i \in [1, m]\}$ and below the others $I \setminus A$, for any choice of the subset A.



The shattering of a set $\{x_1, x_2\}$ of two items wit witnesses t_1 and t_2.

Definition 10.2: Pseudo-Dimension

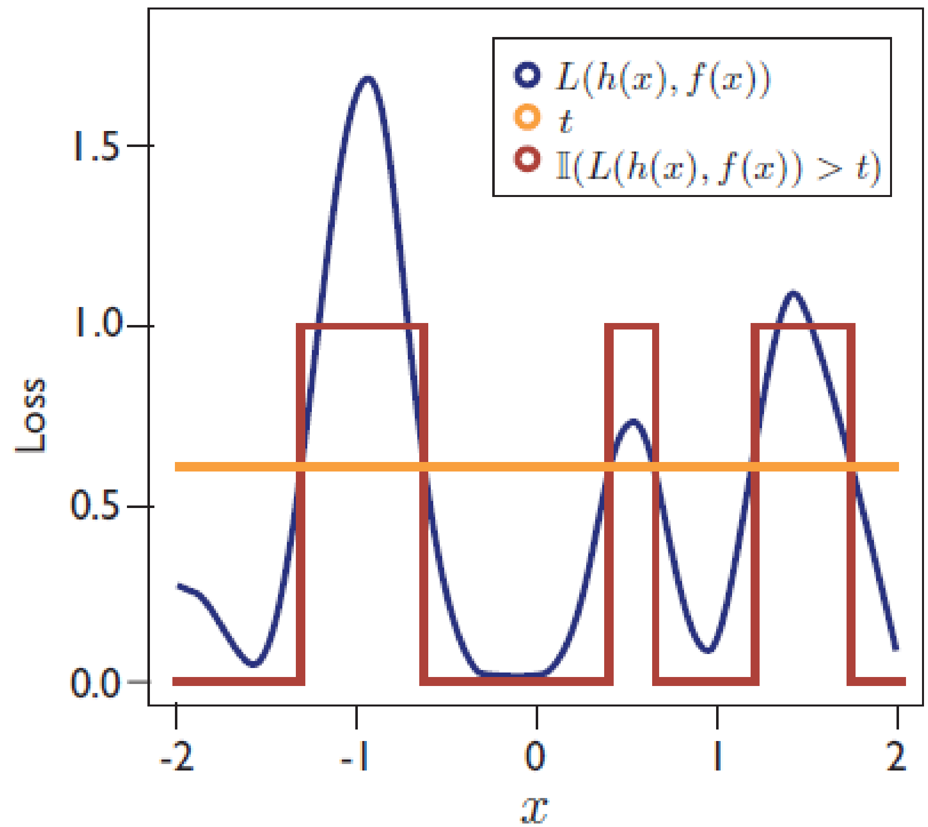
- $\mathscr{I}$: the input space of all possible items.
- $\mathcal{G}$: a family of measurable functions from $\mathscr{I}$ to $\mathbb{R}$.

The pseudo-dimension $\mathrm{Pdim}(\mathcal{G})$ of $\mathcal{G}$ is the size of a largest possible subset of $\mathscr{I}$ shattered by $\mathcal{G}$.

Remarks

- The pseudo-dimension of a family $\mathcal{G}$ of real-valued measurable functions on $\mathscr{I}$ is just the VC-dimension of the corresponding family of binary-valued measurable functions on $\mathscr{I}$ constructed from $g \in \mathcal{G}$ with threshold $t \in \mathbb{R}$:

$$
\begin{aligned}
\mathrm{Pdim}(\mathcal{G}) &= \mathrm{VCdim}(\{\omega \mapsto \mathrm{sgn}(g(\omega) - t) \mid g \in \mathcal{G}, t \in \mathbb{R}\}) \\
&= \mathrm{VCdim}(\{\omega \mapsto 1_{g(\omega)-t>0} \mid g \in \mathcal{G}, t \in \mathbb{R}\}).
\end{aligned}
$$



A function $g : \omega \mapsto L(h(\omega), c(\omega))$ (in blue) defined as the loss of some hypothesis $h \in \mathcal{H}$, and its thresholded version $\omega \mapsto 1_{L(h(\omega),c(\omega))>t}$ (in red) with respect to the threshold t (in yellow).

Pseudo-Dimension of the Family of Affine Functions

Theorem 10.4:

- $\mathscr{I} = \mathbb{R}^N$: the input space.
- $\mathcal{A} = \{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} + b \mid \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$: the family of all affine functions on $\mathbb{R}^N$.

Then $\text{Pdim}(\mathcal{A}) = \text{VCdim}(\text{sgn} \circ \mathcal{A}) = N + 1$.

Pseudo-Dimension of a Vector Space of Real-Values Functions

Theorem 10.5:

- $\mathscr{I}$: the input space.
- $\mathcal{G}$: a finite dimensional vector space of real-valued measurable functions on $\mathscr{I}$.

Then $\mathrm{Pdim}(\mathcal{G}) = \dim(\mathcal{G})$.

Expectation of a Nonnegative Random Variable

- X : a nonnegative random variable on a probability space $(\Omega, \mathcal{F}, P)$.
- $F_X(x)$: the probability distribution function of X.

Then

$$E[X] \triangleq \int_{[0,\infty)} x \; dF_X(x) = \int_{[0,\infty)} P(X > x)dx$$

Proof. Since $g(x) = x$ and $F_X(x)$ are monotone increasing functions on $[0, M]$ for any $M > 0$, by the formula for integration by parts for Riemann-Stieltjes integral[a], we have

---

[a] A.M. Apostol, *Mathematical Analysis,* 2nd edn. Pearson, 1974, Theorem 7.6, page 144.

$$
\begin{aligned}
\int_{[0,M]} x \; dF_X(x) &= (MF_X(M) - 0F_X(0)) - \int_{[0,M]} F_X(x)dx \\
&= \int_{[0,M]} (F_X(M) - F_X(x))dx \\
&= \int_{[0,\infty)} 1_{[0,M]}(x)(F_X(M) - F_X(x))dx.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
E[X] &= \int_{[0,\infty)} x \; dF_X(x) = \lim_{M\to\infty} \int_{[0,M]} x \; dF_X(x) \\
&= \lim_{M\to\infty} \int_{[0,\infty)} 1_{[0,M]}(x)(F_X(M) - F_X(x))dx \\
&= \int_{[0,\infty)} \lim_{M\to\infty} (1_{[0,M]}(x)(F_X(M) - F_X(x)))dx \\
&= \int_{[0,\infty)} (1 - F_X(x))dx
\end{aligned}
$$

by Lebesgue's monotone convergence theorem [a], since

$$1_{[0,M]}(x)(F_X(M) - F_X(x)) \uparrow 1 - F_X(x) \quad \text{as } M \uparrow \infty.$$

$\square$

- It is possible that $E[X] = \infty$ for a nonnegative r.v. X.

[a] W. Rudin, *Principles of Mathematical Analysis*, 3rd end. New York: McGraw-Hill, 1976, Theorem 11.28, pp. 318-319.

Expectation of a Random Variable

- X : a real-valued random variable on a probability space $(\Omega, \mathcal{F}, P)$.
- $X^+ \triangleq \max(0, X)$: the positive part of X.
- $X^- \triangleq \max(0, -X)$: the negative part of X.
- $X = X^+ - X^-$.

The expectation $E[X]$ of X is defined as

$$E[X] \triangleq E[X^+] - E[X^-] = \int_{[0,\infty)} P(X > x)dx - \int_{[0,\infty)} P(X < -x)dx.$$

if either $E[X^+]$ or $E[X^-]$ is finite. Otherwise, $E[X]$ does not exist.

Pseudo-Dimension Regression Generalization Bounds

Theorem 10.6: Let

- c : a fixed but unknown concept in a concept class C of real-valued measurable functions on the input space $\mathscr{I}$ to learn.
- $\mathcal{H}$: a hypothesis set of real-valued measurable functions on $\mathscr{I}$, which may be an infinite set.
- $L : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}^+$: a loss function such that

$$L(h(\omega), c(\omega)) \leq M, \ \forall\, \omega \in \mathscr{I}, \ \forall\, h \in \mathcal{H}.$$

- $\mathcal{G} = \{\omega \mapsto L(h(\omega), c(\omega)) \mid h \in \mathcal{H}\}$: the family of loss functions associated to the hypothesis set $\mathcal{H}$.
  - The range of every loss function in $\mathcal{G}$ contains in $[0, M]$.
- $\mathrm{Pdim}(\mathcal{G}) = d$.

- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the input space $\mathscr{I}$ according to a distribution D with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- $\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} L(h(\omega_i), c(\omega_i)) = \underset{\omega \sim \hat{D}}{E}[L(h(\omega), c(\omega))]$: the empirical loss of the hypothesis h on the sample S, where $\hat{D}$ is the empirical distribution on $\mathscr{I}$ induced by S.
- $R(h) = \underset{\omega \sim D}{E}[L(h(\omega), c(\omega))]$: the expected loss of h.

For any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all h in $\mathcal{H}$:

$$R(h) \leq \hat{R}_S(h) + 2M\sqrt{\frac{2d \ln \frac{em}{d}}{m}} + M\sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

Proof. Since $L(h(\omega), c(\omega))$ is a nonnegative r.v. with values bounded by M, we have

$$
\begin{aligned}
& R(h) - \hat{R}_S(h) \\
= \; & \underset{\omega\sim D}{E}[L(h(\omega), c(\omega))] - \underset{\omega\sim \hat{D}}{E}[L(h(\omega), c(\omega))] \\
= \; & \int_0^M \left(\underset{\omega\sim D}{P}(L(h(\omega), c(\omega)) > t) - \underset{\omega\sim \hat{D}}{P}(L(h(\omega), c(\omega)) > t) \right) dt \\
\leq \; & M \sup_{t\in[0,M]} \left(\underset{\omega\sim D}{P}(L(h(\omega), c(\omega)) > t) - \underset{\omega\sim \hat{D}}{P}(L(h(\omega), c(\omega)) > t) \right) \\
\leq \; & M \sup_{t\in\mathbb{R}} \left(\underset{\omega\sim D}{E}[1_{L(h(\omega),c(\omega))>t}] - \underset{\omega\sim \hat{D}}{E}[1_{L(h(\omega),c(\omega))>t}] \right) \\
= \; & M \sup_{t\in\mathbb{R}} \left(\underset{\omega\sim D}{E}[1_{L(h(\omega),c(\omega))>t}] - \frac{1}{m}\sum_{i=1}^{m} 1_{L(h(\omega_i),c(\omega_i))>t} \right)
\end{aligned}
$$

for all $h \in \mathcal{H}$ so that

$$
\begin{aligned}
& \sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \\
\leq \quad & M \sup_{h \in \mathcal{H}, t \in \mathbb{R}} \left(\underset{\omega \sim D}{E}[1_{L(h(\omega), c(\omega)) > t}] - \frac{1}{m} \sum_{i=1}^{m} 1_{L(h(\omega_i), c(\omega_i)) > t} \right) \\
= \quad & M \sup_{g \in \mathcal{G}, t \in \mathbb{R}} \left(\underset{\omega \sim D}{E}[1_{g(\omega) - t > 0}] - \frac{1}{m} \sum_{i=1}^{m} 1_{g(\omega_i) - t > 0} \right) \\
= \quad & M \sup_{g_t \in \tilde{\mathcal{G}}} \left(\underset{\omega \sim D}{E}[g_t(\omega)] - \frac{1}{m} \sum_{i=1}^{m} g_t(\omega_i) \right),
\end{aligned}
$$

where

- $g_t : \omega \mapsto 1_{g(\omega) - t > 0}$: the classifier corresponding to the loss function $g(\omega) = L(h(\omega), c(\omega))$ in $\mathcal{G}$ with threshold t, whose range is $\{0, 1\}$.
- $\tilde{\mathcal{G}} = \{g_t \mid g \in \mathcal{G}, t \in \mathbb{R}\}$.

By applying the Rademacher complexity bound (Theorem 3.1) to the family $\tilde{\mathcal{G}}$, we have: for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sup_{g_t \in \tilde{\mathcal{G}}} \left(\underset{\omega \sim D}{E}[g_t(\omega)] - \frac{1}{m} \sum_{i=1}^{m} g_t(\omega_i) \right) \quad \leq \quad 2\mathfrak{R}_m(\tilde{\mathcal{G}}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

which implies: for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \leq 2M\mathfrak{R}_m(\tilde{\mathcal{G}}) + M\sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

Next applying Corollary 3.1 and Corollary 3.3 to the family $\tilde{\mathcal{G}}$ of binary classifiers, we have

$$\mathfrak{R}_m(\tilde{\mathcal{G}}) \leq \sqrt{\frac{2 \ln \Pi_{\tilde{\mathcal{G}}}(m)}{m}} \text{ and } \Pi_{\tilde{\mathcal{G}}}(m) \leq \left(\frac{em}{d} \right)^d$$

where $\Pi_{\tilde{\mathcal{G}}}(m)$ is the growth function of the family $\tilde{\mathcal{G}}$ and $d = \mathrm{VCdim}(\tilde{\mathcal{G}})$. Since $\mathrm{VCdim}(\tilde{\mathcal{G}}) = \mathrm{Pdim}(\mathcal{G})$, we have:

for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \leq 2M \sqrt{\frac{2d \ln \frac{em}{d}}{m}} + M \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}.$$

$\square$

The Contents of This Lecture

- Generalization bounds
- Linear regression
- Kernel ridge regression
- Support vector regression
- Lasso

Motivations

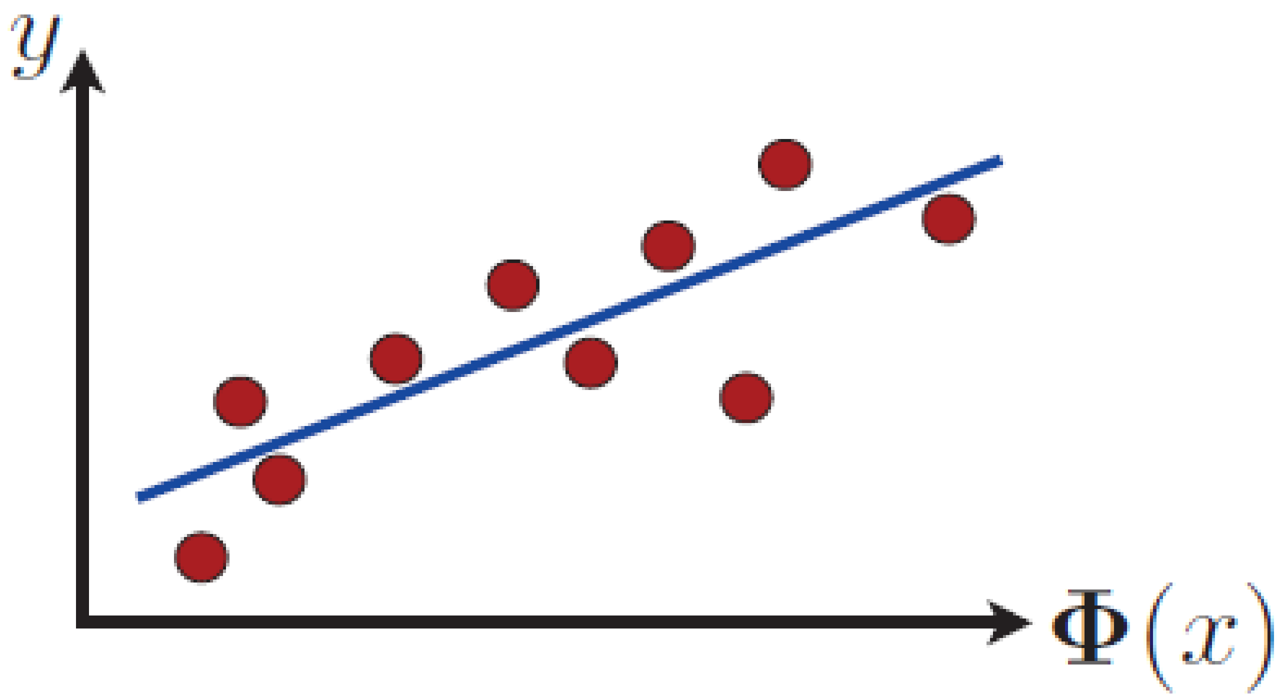
- The generalization results in above show that, for the same empirical error, hypothesis sets with smaller complexity, measured in terms of the Rademacher complexity or in terms of pseudo-dimension, benefit from better generalization guarantees.
- One family of functions with relatively small complexity is that of linear hypotheses. In the rest of this lecture, we describe and analyze several algorithms based on that hypothesis set.

Linear Regression Problem

- $\mathscr{I}$: the input space of all possible items, associated with a probability space $(\mathscr{I}, \mathcal{F}, D)$.
- $\mathscr{Y}' = \mathscr{Y} \subseteq \mathbb{R}$: the output and label spaces, which typically are $[-M, M]$ for some $M > 0$ or $\mathbb{R}$.
- $c : \mathscr{I} \to \mathscr{Y}$: a fixed but unknown target concept in a concept class $\mathcal{C}$ of $\mathscr{Y}$-valued measurable function on $\mathscr{I}$.
- $\Phi : \mathscr{I} \to \mathbb{R}^N$: a feature mapping from the input space $\mathscr{I}$ to the N-dimensional feature space $\mathbb{R}^N$.
- $\mathcal{H} = \{\omega \mapsto \mathbf{w} \cdot \Phi(\omega) + b \mid \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$: the hypothesis set of all affine functions in the feature space $\mathbb{R}^N$.
- $L(y', y) = (y' - y)^2$: the squared-error loss function.

- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the input space according to D, with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- Empirical Risk Minimization Problem : find a hypothesis $h_S : \mathscr{I} \to \mathscr{Y}$ in $\mathcal{H}$ with the smallest empirical mean squared error w.r.t. the target concept c,

$$\min_{\mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}} F(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \left(\mathbf{w} \cdot \Phi(\omega_i) + b - c(\omega_i)\right)^2 .$$



Linear regression with $N = 1$.

Minimizing the Object Function of Linear Regression

$$F(\mathbf{W}) = \frac{1}{m}\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|^2$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} \Phi(\omega_1) & \Phi(\omega_2) & \dots & \Phi(\omega_m) \\ 1 & 1 & \dots & 1 \end{bmatrix}, \ \text{ and } \mathbf{Y} = \begin{bmatrix} c(\omega_1) \\ c(\omega_2) \\ \vdots \\ c(\omega_m) \end{bmatrix}.$$

- $F(\mathbf{W})$ is a quadratic function of $\mathbf{W}$.
- $F(\mathbf{W})$ is minimized if and only if $\nabla F(\mathbf{W}) = \mathbf{0}$ if and only if $\frac{2}{m}\mathbf{X}(\mathbf{X}^T\mathbf{W} - \mathbf{Y}) = \mathbf{0}$ if and only if $\mathbf{X}\mathbf{X}^T\mathbf{W} = \mathbf{X}\mathbf{Y}$.

Solution of the Minimization of $F(\mathbf{W})$

- If the $(N+1) \times (N+1)$ matrix $\mathbf{X}$ is invertible, then $F(\mathbf{W})$ is minimized at a unique point $\mathbf{W}^{LR} = (\mathbf{X}^T)^{-1}\mathbf{Y}$.
- If $\mathbf{X}$ is not invertible, then $F(\mathbf{W})$ is minimized at infinitely many points, which can be expressed by the Moore-Penrose pseudoinverse $(\mathbf{X}^T)^+$ of $\mathbf{X}^T$ as follows:
  - $\mathbf{X} = \mathbf{U_X \Sigma_X V_X^T}$: a singular value decomposition of $\mathbf{X}$.
  - $\mathbf{X}^T = \mathbf{V_X \Sigma_X U_X^T}$: a singular value decomposition of $\mathbf{X}^T$.
  - $\mathbf{XX}^T = \mathbf{U_X \Sigma_X V_X^T V_X \Sigma_X U_X^T} = \mathbf{U_X \Sigma_X^2 U_X^T}$.
  - $\mathbf{XX}^T\mathbf{W} = \mathbf{XY} \Leftrightarrow \mathbf{U_X \Sigma_X^2 U_X^T W} = \mathbf{U_X \Sigma_X V_X^T Y} \Leftrightarrow$ $\mathbf{\Sigma_X^2 U_X^T W} = \mathbf{\Sigma_X V_X^T Y}$ by multiplying both sides with $\mathbf{U_X^T}$ $\Leftrightarrow \mathbf{U_X^T W} = \mathbf{\Sigma_X^{-1} V_X^T Y}$ by multiplying both sides with $\mathbf{\Sigma_X^{-2}}$.

- Both $\mathbf{W}$ and $\mathbf{W}'$ are solutions of the least-square equation if and only if $\mathbf{U}_{\mathbf{X}}^T(\mathbf{W} - \mathbf{W}') = \mathbf{0}$ if and only if $\mathbf{U}_{\mathbf{X}}\mathbf{U}_{\mathbf{X}}^T(\mathbf{W} - \mathbf{W}') = \mathbf{0}$ since $\mathbf{U}_{\mathbf{X}}$ has full rank if and only if $\mathbf{W} - \mathbf{W}' = (\mathbf{I} - \mathbf{U}_{\mathbf{X}}\mathbf{U}_{\mathbf{X}}^T)\mathbf{W}_0$, where $\mathbf{U}_{\mathbf{X}}\mathbf{U}_{\mathbf{X}}^T$ is the orthogonal projection of the column space of $\mathbf{U}_{\mathbf{X}}$ and $\mathbf{W}_0$ is an arbitrary vector in $\mathbb{R}^{N+1}$.
- $(\mathbf{X}^T)^+ = \mathbf{U}_{\mathbf{X}}\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{V}_{\mathbf{X}}^T$: the Moore-Penrose pseudoinverse of $\mathbf{X}$.
- Since $\mathbf{U}_{\mathbf{X}}^T((\mathbf{X}^T)^+\mathbf{Y}) = \mathbf{U}_{\mathbf{X}}^T\mathbf{U}_{\mathbf{X}}\mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{V}_{\mathbf{X}}^T\mathbf{Y} = \mathbf{\Sigma}_{\mathbf{X}}^{-1}\mathbf{V}_{\mathbf{X}}^T\mathbf{Y}$, $(\mathbf{X}^T)^+\mathbf{Y}$ is a solution of the least-square equation.
- Now $\mathbf{W}^{LR} = (\mathbf{X}^T)^+\mathbf{Y} + (\mathbf{I} - \mathbf{U}_{\mathbf{X}}\mathbf{U}_{\mathbf{X}}^T)\mathbf{W}_0$ is a general solution of the least-square equation, where $\mathbf{W}_0$ is an arbitrary vector in $\mathbb{R}^{N+1}$.

- Since the solution $(\mathbf{X}^T)^+\mathbf{Y} = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}_\mathbf{X}^{-1}\mathbf{V}_\mathbf{X}^T\mathbf{Y}$ is in the range of the orthogonal projection $\mathbf{U}_\mathbf{X}\mathbf{U}_\mathbf{X}^T$, it is orthogonal to $(\mathbf{I} - \mathbf{U}_\mathbf{X}\mathbf{U}_\mathbf{X}^T)\mathbf{W}_0$ and then has the minimum length among all solutions and is often preferred for that reason.

- The solution to minimize $F(\mathbf{W})$ is

$$\mathbf{W}^{LR} = \begin{cases} (\mathbf{X}^T)^{-1}\mathbf{Y}, & \text{if } \mathbf{X} \text{ is invertible,} \\ (\mathbf{X}^T)^{+}\mathbf{Y}, & \text{otherwise.} \end{cases}$$

Computational Complexity of Linear Regression

- The cost of computing the inverse or the pseudoinverse of $\mathbf{X}^T$ is in $O(N^{2+\omega})$ with $\omega = 0.376$ by a method such as that of Coppersmith and Winograd.
- The multiplication of $(\mathbf{X}^T)^{-1}$ or $(\mathbf{X}^T)^{+}$ with $\mathbf{Y}$ takes $O(mN)$.
- Therefore, the overall complexity of computing the solution $\mathbf{W}^{LR}$ is in $O(mN + N^{2+\omega})$.
- Thus, when the dimension of the feature space N is not too large, the solution can be computed efficiently.

Remarks

- While linear regression is simple and admits a straightforward implementation, it does not benefit from a strong generalization guarantee, since it is limited to minimizing the empirical error without controlling the norm of the weight vector and without any other regularization.
- Its performance is also typically poor in most applications.
- The next sections describe algorithms with both better theoretical guarantees and improved performance in practice.

The Contents of This Lecture

- Generalization bounds
- Linear regression
- Kernel ridge regression
- Support vector regression
- Lasso

Mean Square Error Bound for Kernel-Based Hypotheses

Theorem 10.7: Let

- $\mathscr{I}$: the input space of all possible items ω, associated with a probability space $(\mathscr{I}, \mathcal{F}, D)$.
- $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a PDS kernel over the input space $\mathscr{I}$ with $K(\omega, \omega) \leq r^2$ for all $\omega \in \mathscr{I}$.
- $\Phi : \mathscr{I} \to \mathscr{F}$: a feature mapping associated to the PDS kernel K from the input space $\mathscr{I}$ to a feature space $\mathscr{F}$ such that for all $\omega, \omega' \in \mathscr{I}$, $\langle \Phi(\omega), \Phi(\omega') \rangle_{\mathscr{F}} = K(\omega, \omega')$.
- $\mathcal{H} = \{\omega \mapsto \langle f, \Phi(\omega) \rangle_{\mathscr{F}} + b \mid \|f\|_{\mathscr{F}} \leq \Lambda, |b| \leq r\Lambda\}$: a kernel-based hypothesis set.

- c : a fixed but unknown concept in a concept class C of real-valued measurable functions on the input space $\mathscr{I}$ to learn with $|c(\omega)| \leq r\Lambda$ for all $\omega \in \mathscr{I}$.
- $L(y', y) = (y' - y)^2$: the squared error loss function.
- $S = (\omega_1, \ldots, \omega_m)$: a labeled sample of size m drawn i.i.d. according to the distribution D with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- $R(h) = \underset{\omega \sim D}{E}[(h(\omega) - c(\omega))^2]$: the expected mean squared error.
- $\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} (h(\omega_i) - c(\omega_i))^2$: the empirical mean squared error.

Then for any $\delta > 0$, with probability at least $1-\delta$, each of the following regression generalization bounds holds for all h in $\mathcal{H}$:

$$
\begin{aligned}
R(h) &\leq \hat{R}_S(h) + \frac{24r^2\Lambda^2}{\sqrt{m}}\left(1 + \frac{3}{8}\sqrt{\frac{\ln\frac{1}{\delta}}{2}}\right), \\
R(h) &\leq \hat{R}_S(h) + \frac{12r^2\Lambda^2}{\sqrt{m}}\left(1 + \sqrt{\frac{\operatorname{tr}(\mathbf{K})}{mr^2}} + \frac{9}{4}\sqrt{\frac{\ln\frac{2}{\delta}}{2}}\right),
\end{aligned}
$$

where $\mathbf{K} = [K(\omega_i, \omega_j)]$ is the $m \times m$ kernel matrix associated to the kernel K and the sample S.

Proof. For any $h \in \mathcal{H}$, we have

$$|h(\omega)| = |\langle f, \Phi(\omega)\rangle_{\mathscr{F}} + b| \leq \|f\|_{\mathscr{F}} \|\Phi(\omega)\|_{\mathscr{F}} + |b| \leq 2\Lambda r \ \forall\ \omega \in \mathscr{I},$$

by Cauchy-Schwartz inequality, so that

$$|h(\omega) - c(\omega)| \leq 3r\Lambda \ \forall\ \omega \in \mathscr{I}.$$

By the Rademacher complexity regression generalization bounds (Theorem 10.3) with $p = 2$ and $M = 3r\Lambda$, we have: for any $\delta > 0$, with probability at least $1 - \delta$, each of the following holds for all h in $\mathcal{H}$:

$$\begin{aligned} R(h) &\leq \hat{R}_S(h) + 12r\Lambda\mathfrak{R}_m(\mathcal{H}) + 9r^2\Lambda^2\sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \\ R(h) &\leq \hat{R}_S(h) + 12r\Lambda\hat{\mathfrak{R}}_S(\mathcal{H}) + 27r^2\Lambda^2\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}. \end{aligned}$$

Also from Theorem 5.5, the empirical Rademacher complexity of the kernel-based hypothesis set $\mathcal{H}$ w.r.t. the sample S can be bounded by

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{\Lambda\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + \frac{r\Lambda}{\sqrt{m}} \leq \frac{2r\Lambda}{\sqrt{m}},$$

where $\mathbf{K} = [K(\omega_i, \omega_j)]$ is the $m \times m$ kernel matrix associated to the kernel K and the sample S. And by averaging over all samples S, we have

$$\mathfrak{R}_m(\mathcal{H}) \leq \frac{2r\Lambda}{\sqrt{m}}.$$

This completes the proof. $\square$

Remarks

- The first bound in Theorem 10.7 has the form

$$R(h) \leq \hat{R}_S(h) + \lambda\Lambda^2$$

with $\lambda = \frac{24r^2}{\sqrt{m}}\left(1 + \frac{3}{8}\sqrt{\frac{\ln\frac{1}{\delta}}{2}}\right) = O(\frac{1}{\sqrt{m}})$.

- Ridge regression is defined by the minimization of an objective function that has precisely this form and thus is directly motivated by the theoretical analysis just presented.

Kernel Ridge Regression Problem

- c: a fixed but unknown target concept in a concept class $\mathcal{C}$ of real-valued measurable function on the input space $\mathscr{I}$.
- $\Phi : \mathscr{I} \to \mathbb{R}^N$: a feature mapping from the input space $\mathscr{I}$ to the N-dimensional feature space $\mathbb{R}^N$.
- $\mathcal{H} = \{\omega \mapsto \mathbf{w} \cdot \Phi(\omega) \mid \mathbf{w} \in \mathbb{R}^N\}$: the hypothesis set of all linear functions in the feature space $\mathbb{R}^N$.
- $L(y', y) = (y' - y)^2$: the squared-error loss function.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the input space according to D, with labels $(c(\omega_1), \ldots, c(\omega_m))$.

- Problem : find a hypothesis $h_S : \mathscr{I} \to \mathscr{Y}$ in $\mathcal{H}$, i.e., a weight vector $\mathbf{w} \in \mathbb{R}^N$, which minimizes the following object function

$$F(\mathbf{w}) = \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^{m} \left(\mathbf{w} \cdot \Phi(\omega_i) - c(\omega_i)\right)^2 .$$

Remarks

- The parameter λ is a positive parameter determining the trade-off between the regularization term $\|\mathbf{w}\|^2$ and the empirical mean squared error.
- Except for the shift $b = 0$ in the second term, the objective function of the kernel ridge regression differs from that of linear regression only by the first term, which controls the norm of the weight vector $\mathbf{w}$.

Minimization of the Kernel Ridge Regression Object Function

The object function of the kernel ridge regression problem can be written as:

$$F(\mathbf{w}) = \lambda\|\mathbf{w}\|^2 + \|\mathbf{X}^T\mathbf{w} - \mathbf{Y}\|^2$$

where $\mathbf{X} = [\Phi(\omega_1), \Phi(\omega_2), \ldots, \Phi(\omega_m)]$ is an $N \times m$ matrix and $\mathbf{Y} = [c(\omega_1), c(\omega_2), \ldots, c(\omega_m)]^T$ is an $m \times 1$ matrix.

- $F(\mathbf{w})$ is a quadratic function of $\mathbf{w}$.
- $F(\mathbf{w})$ is minimized if and only if $\nabla F(\mathbf{w}) = \mathbf{0}$ if and only if $2\lambda\mathbf{w} + 2\mathbf{X}(\mathbf{X}^T\mathbf{w} - \mathbf{Y}) = \mathbf{0}$ if and only if $(\lambda\mathbf{I} + \mathbf{X}\mathbf{X}^T)\mathbf{w} = \mathbf{X}\mathbf{Y}$.
- Since $\lambda > 0$, the eigenvalues of the symmetric matrix $(\lambda\mathbf{I} + \mathbf{X}\mathbf{X}^T)$ are all positive and then $(\lambda\mathbf{I} + \mathbf{X}\mathbf{X}^T)$ is invertible.
- The unique solution to minimize $F(\mathbf{w})$ is

$$\mathbf{w}^{KRR} = (\lambda\mathbf{I} + \mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}.$$

Alternative Formulations of KRR Primal Problem

$$\begin{aligned} \text{Minimize} \quad & F(\mathbf{w}) = \sum_{i=1}^{m} \left(\mathbf{w} \cdot \Phi(\omega_i) - c(\omega_i)\right)^2 \\ \text{Subject to} \quad & \|\mathbf{w}\|^2 - \Lambda^2 \leq 0 \\ & \mathbf{w} \in \mathbb{R}^N. \end{aligned}$$

or equivalently, using slack variables $\eta_i = c(\omega_i) - \mathbf{w} \cdot \Phi(\omega_i)$, $i \in [1, m]$,

$$\begin{aligned} \text{Minimize} \quad & F(\mathbf{w}, \eta) = \frac{1}{2} \sum_{i=1}^{m} \eta_i^2 \\ \text{Subject to} \quad & \frac{1}{2}\|\mathbf{w}\|^2 - \Lambda^2 \leq 0 \\ & c(\omega_i) - \mathbf{w} \cdot \Phi(\omega_i) - \eta_i = 0, \ i \in [1, m] \\ & \mathbf{w} \in \mathbb{R}^N, \ \xi \in \mathbb{R}^m. \end{aligned}$$

Qualification of the Primal Problem

- The object function $F(\mathbf{w}, \eta) = \frac{1}{2} \sum_{i=1}^{m} \eta_i^2$ is infinitely differentiable and convex so that it is pseudoconvex at any feasible point.
- The inequality constraint function $g(\mathbf{w}, \eta) = \frac{1}{2}\|\mathbf{w}\|^2 - \Lambda^2$ is infinitely differentiable and convex so that it is pseudoconvex at any feasible point.
- The equality constraint functions $h_i(\mathbf{w}, \eta) = c(\mathbf{x}_i) - \mathbf{w} \cdot \Phi(\omega_i) - \eta_i$, $1 \leq i \leq m$, are affine functions so that they are infinitely differentiable, convex and concave and then quasiconvex and quasiconcave at any feasible point.
- $\nabla F = \begin{bmatrix} \mathbf{0} \\ \eta \end{bmatrix}$, $\nabla g = \begin{bmatrix} \mathbf{w} \\ \mathbf{0} \end{bmatrix}$, and $\nabla h_i = \begin{bmatrix} -\Phi(\omega_i) \\ -\mathbf{e}_i \end{bmatrix}$.

- The Kuhn-Tucker necessary conditions are:

$$\nabla F + \lambda \nabla g + \sum_{i=1}^{m} \mu_i \nabla h_i = \mathbf{0}$$
$$\Leftrightarrow \lambda \mathbf{w} = \sum_{i=1}^{m} \mu_i \Phi(\omega_i), \eta_i = \mu_i, i \in [1, m]$$
$$\lambda g(\mathbf{w}, \eta) = 0$$
$$\lambda \geq 0.$$

- Any feasible point $(\mathbf{w}, \eta)$ which satisfies the Kuhn-Tucker necessary conditions in above is a global minimum solution.
- If $\lambda > 0$, the weight vector solution $\mathbf{w}^{KRR}$ is a linear combination of the training feature vectors $\Phi(\omega_1), \ldots, \Phi(\omega_m)$,

$$\mathbf{w}^{KRR} = \frac{1}{\lambda} \sum_{i=1}^{m} \mu_i \Phi(\omega_i),$$

  and has $\|\mathbf{w}^{KRR}\|^2 = 2\Lambda^2$.
  - If $\Phi(\omega_1), \ldots, \Phi(\omega_m)$ are linearly independent, then $\lambda > 0$.

- If $\lambda = 0$, then we must have $\mathbf{X}\eta^{KRR} = \mathbf{0}$, i.e., $\Phi(\omega_1), \ldots, \Phi(\omega_m)$ are linearly dependent, and $\mathbf{Y} - \mathbf{X}^T\mathbf{w}^{KRR} - \eta^{KRR} = \mathbf{0}$. Thus we have infinitely many solutions

$$\mathbf{w}^{KRR} = (\mathbf{X}^T)^+\mathbf{Y} + (\mathbf{I} - \mathbf{U}_\mathbf{X}\mathbf{U}_\mathbf{X}^T)\mathbf{w}_0,$$

where $\mathbf{U}_\mathbf{X}\mathbf{U}_\mathbf{X}^T$ is a projection and $\mathbf{w}_0$ is any vector in $\mathbb{R}^N$, and among them,

$$\mathbf{w}^{KRR} = (\mathbf{X}^T)^+\mathbf{Y}$$

has the minimum $\|\mathbf{w}^{KRR}\|^2$ since $(\mathbf{X}^T)^+\mathbf{Y}$ is in the range of the projection $\mathbf{U}_\mathbf{X}\mathbf{U}_\mathbf{X}^T$.

  - When $\lambda = 0$ (and then $\mathbf{X}\mathbf{X}^T$ not invertible), $\mathbf{w}^{KRR}$ is the same as that obtained by linear regression.

The Returned Hypothesis h_S^{KRR} by KRR

With $\lambda > 0$, the returned hypothesis h_S^{KRR} by KRR is

$$h_S^{KRR}(\omega) = \mathbf{w}^{KRR} \cdot \Phi(\omega) = \frac{1}{\lambda}\sum_{i=1}^{m} \mu_i \Phi(\omega_i) \cdot \Phi(\omega) = \frac{1}{\lambda}\sum_{i=1}^{m} \mu_i K(\omega_i, \omega),$$

where

$$K(\omega_i, \omega) \triangleq \Phi(\omega_i) \cdot \Phi(\omega)$$

is the PDS kernel associated with the feature mapping Φ.

- With this formulation, KRR can be extended to an arbitrary PDS kernel K over the input space $\mathscr{I}$, where a Hilbert space $\mathbb{H}$ and a feature mapping $\Phi : \mathscr{I} \to \mathbb{H}$ can be associated.
- We will use the Lagrangian dual problem to solve the Lagrange multipliers μ_i's.

Lagrangian Dual Function for KRR

- $X = \mathbb{R}^N \times \mathbb{R}^m$: a nonempty open convex set.
- Lagrangian function: for all $\mathbf{w} \in \mathbb{R}^N, \eta \in \mathbb{R}^m$ and $\lambda \in \mathbb{R}, \mu \in \mathbb{R}^m$,

$$\begin{aligned} & L(\mathbf{w}, \eta, \lambda, \mu) \\ = \quad & F(\mathbf{w}, \eta) + \lambda g(\mathbf{w}, \eta) + \sum_{i=1}^{m} \mu_i h_i(\mathbf{w}, \eta) \\ = \quad & \frac{1}{2}\|\eta\|^2 + \lambda(\frac{1}{2}\|\mathbf{w}\|^2 - \Lambda^2) + \sum_{i=1}^{m} \mu_i (c(\omega_i) - \mathbf{w} \cdot \Phi(\omega_i) - \eta_i). \end{aligned}$$

- For any fixed $\lambda \in \mathbb{R}, \mu \in \mathbb{R}^m$, the gradient ∇L of the

Lagrangian function w.r.t. $(\mathbf{w}, \eta)$ is

$$\begin{aligned}
\nabla L &= \nabla F + \lambda \nabla g + \sum_{i=1}^{m} \mu_i \nabla h_i \\
&= \begin{bmatrix} \mathbf{0} \\ \eta \end{bmatrix} + \begin{bmatrix} \lambda \mathbf{w} \\ \mathbf{0} \end{bmatrix} - \sum_{i=1}^{m} \lambda_i \begin{bmatrix} \Phi(\omega_i) \\ \mathbf{e}_i \end{bmatrix}
\end{aligned}$$

and the Hessian matrix is

$$\mathbf{H} = \begin{bmatrix} \lambda \mathbf{I}_{N \times N} & \mathbf{0}_{N \times m} \\ \mathbf{0}_{m \times N} & \mathbf{I}_{m \times m} \end{bmatrix}$$

which is positive semi-definite.

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

$\nabla L(\hat{\mathbf{w}}, \hat{\eta}, \lambda, \mu) = \mathbf{0}$ if and only if

$$\lambda \hat{\mathbf{w}} = \sum_{i=1}^{m} \mu_i \Phi(\omega_i) \text{ and } \hat{\eta}_i = \mu_i,\ i \in [1, m].$$

– Note that for $\lambda = 0$ and any fixed $\mu \in \mathbb{R}^m$, $\sum_{i=1}^{m} \mu_i \Phi(\omega_i) \neq \mathbf{0}$ if and only if the infimum of the Lagrangian function $L(\mathbf{w}, \eta, \lambda, \mu)$ over X is $-\infty$.

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

$$
\begin{aligned}
& \theta(\lambda, \mu) \\
= \quad & \inf_{(\mathbf{w},\eta)\in X} L(\mathbf{w}, \eta, \lambda, \mu) \\
= \quad & \begin{cases}
\frac{1}{2}\|\hat{\eta}\|^2 + \lambda(\frac{1}{2}\|\hat{\mathbf{w}}\|^2 - \Lambda^2) \\
+\sum_{i=1}^m \mu_i(c(\omega_i) - \hat{\mathbf{w}} \cdot \Phi(\omega_i) - \hat{\eta}_i), \ \text{if } \lambda \neq 0, \\
\frac{1}{2}\|\hat{\eta}\|^2 + \sum_{i=1}^m \mu_i(c(\omega_i) - \hat{\eta}_i), \ \text{if } \lambda = 0, \sum_{i=1}^m \mu_i \Phi(\omega_i) = \mathbf{0} \\
-\infty, \ \text{otherwise}
\end{cases} \\
= \quad & \begin{cases}
\sum_{i=1}^m \mu_i c(\omega_i) - \frac{1}{2}\sum_{i=1}^m \mu_i^2 - \frac{1}{2\lambda}\sum_{i,j=1}^m \mu_i\mu_j \Phi(\omega_i)\cdot\Phi(\omega_j) \\
\qquad -\lambda\Lambda^2, \ \text{if } \lambda \neq 0, \\
\sum_{i=1}^m \mu_i c(\omega_i) - \frac{1}{2}\sum_{i=1}^m \mu_i^2, \ \text{if } \lambda = 0, \sum_{i=1}^m \mu_i \Phi(\omega_i) = \mathbf{0} \\
-\infty, \ \text{otherwise}
\end{cases}
\end{aligned}
$$

Lagrangian Dual Problem for KRR

$$
\begin{array}{ll}
\text{Maximize} & \theta(\lambda, \mu) \\
\text{Subject to} & \lambda \geq 0 \\
 & (\lambda, \mu) \in \mathbb{R} \times \mathbb{R}^m,
\end{array}
$$

where the Lagrangian dual function $\theta(\lambda, \mu)$ is

$$
\begin{aligned}
&\theta(\lambda, \mu) \\
= &\left\{
\begin{array}{l}
\sum_{i=1}^m \mu_i c(\omega_i) - \frac{1}{2} \sum_{i=1}^m \mu_i^2 - \frac{1}{2\lambda} \sum_{i,j=1}^m \mu_i \mu_j \Phi(\omega_i) \cdot \Phi(\omega_j) \\
\qquad\quad -\lambda \Lambda^2, \ \text{if } \lambda \neq 0, \\
\sum_{i=1}^m \mu_i c(\omega_i) - \frac{1}{2} \sum_{i=1}^m \mu_i^2, \ \text{if } \lambda = 0, \sum_{i=1}^m \mu_i \Phi(\omega_i) = \mathbf{0} \\
-\infty, \ \text{otherwise.}
\end{array}
\right.
\end{aligned}
$$

Lagrangian Dual Problem for KRR with a Fixed $\lambda > 0$

$$
\begin{aligned}
\text{Maximize} \quad & \theta_\lambda(\mu) = \sum_{i=1}^m \mu_i c(\omega_i) - \tfrac{1}{2}\sum_{i=1}^m \mu_i^2 \\
& -\tfrac{1}{2\lambda}\sum_{i,j=1}^m \mu_i \mu_j K(\omega_i, \omega_j) - \lambda \Lambda^2 \\
\text{Subject to} \quad & \mu \in \mathbb{R}^m,
\end{aligned}
$$

where $K(\omega, \omega') = \Phi(\omega) \cdot \Phi(\omega') \; \forall\, \omega, \omega' \in \mathscr{I}$ is the PDS kernel associated with the feature mapping Φ, or in vector form,

$$
\begin{aligned}
\text{Maximize} \quad & \theta_\lambda(\mu) = \mu^T \mathbf{Y} - \tfrac{1}{2}\mu^T \mu - \tfrac{1}{2\lambda}\mu^T \mathbf{K} \mu - \lambda \Lambda^2 \\
\text{Subject to} \quad & \mu \in \mathbb{R}^m,
\end{aligned}
$$

where $\mathbf{Y} = [c(\omega_1), \ldots, c(\omega_m)]^T$ and $\mathbf{K} = [K(\omega_i, \omega_j)]$.

- $\theta_\lambda(\mu)$ is maximized if and only if $\nabla \theta_\lambda(\mu) = \mathbf{0}$ if and only if $\mathbf{Y} - \mu - \frac{1}{\lambda}\mathbf{K}\mu = \mathbf{0}$ if and only if $\mu^{KRR} = \lambda(\lambda \mathbf{I} + \mathbf{K})^{-1}\mathbf{Y}$.

- When the feature space $\mathbb{H}$ is $\mathbb{R}^N$, we have $\mathbf{K} = \mathbf{X}^T\mathbf{X}$ and from the Kuhn-Tucker necessary conditions of the primal problem, the optimal weight vector is

$$\mathbf{w}^{KRR} = \frac{1}{\lambda}\mathbf{X}\mu^{KRR} = \mathbf{X}(\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{Y} = (\lambda\mathbf{I} + \mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{Y}$$

  which is the same as obtained from the original KRR problem.

- The returned hypothesis is

$$h_S^{KRR}(\omega) = \frac{1}{\lambda}\sum_{i=1}^{m}\mu_i^{KRR}K(\omega_i,\omega) = \sum_{i=1}^{m}((\lambda\mathbf{I}+\mathbf{K})^{-1}\mathbf{Y})_i K(\omega_i,\omega).$$

- $\max_{\mu\in\mathbb{R}^m}\theta_\lambda(\mu) = \theta_\lambda(\mu^{KRR}) = \frac{\lambda}{2}\mathbf{Y}^T(\lambda\mathbf{I}+\mathbf{K})^{-1}\mathbf{Y} - \lambda\Lambda^2.$

A Useful Lemma

Lemma 10.1: For any $\lambda > 0$ and any $m \times n$ matrix $\mathbf{X}$, we have

$$\mathbf{X}(\lambda \mathbf{I}_{n\times n} + \mathbf{X}^T\mathbf{X})^{-1} = (\lambda \mathbf{I}_{m\times m} + \mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}.$$

Proof. Observe the following identity,

$$\mathbf{X}(\lambda \mathbf{I}_{n\times n} + \mathbf{X}^T\mathbf{X}) = (\lambda \mathbf{I}_{m\times m} + \mathbf{X}\mathbf{X}^T)\mathbf{X},$$

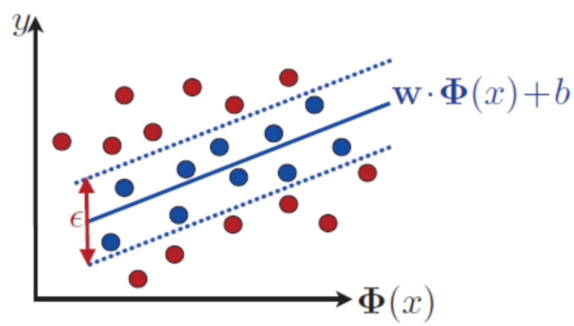
and note that both $(\lambda \mathbf{I}_{n\times n} + \mathbf{X}^T\mathbf{X})$ and $(\lambda \mathbf{I}_{m\times m} + \mathbf{X}\mathbf{X}^T)$ are invertible since $\lambda > 0$. $\square$

The Contents of This Lecture

- Generalization bounds
- Linear regression
- Kernel ridge regression
- Support vector regression
- Lasso

Main Idea of Support Vector Regression

- To fit a tube of width $\epsilon > 0$ along a hyperplane to the data.



SVR attempts to fit a "tube" with width ϵ to the data.

- As in binary classification, this defines two sets of points: those falling inside the tube, which are ϵ-close to the function predicted and thus not penalized, and those falling outside, which are penalized based on their distance to the predicted function, in a way that is similar to the penalization used by SVMs in classification.

Support Vector Regression (SVR) Problem

- c: a fixed but unknown target concept in a concept class $\mathcal{C}$ of real-valued measurable function on the input space $\mathscr{I}$.
- $\Phi : \mathscr{I} \to \mathbb{R}^N$: a feature mapping from the input space $\mathscr{I}$ to the N-dimensional feature space $\mathbb{R}^N$.
- $\mathcal{H} = \{\omega \mapsto \mathbf{w} \cdot \Phi(\omega) + b \mid \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$: the hypothesis set of all affine functions in the feature space $\mathbb{R}^N$.
- $L(y', y) = |y' - y|_\epsilon \triangleq \max(0, |y' - y| - \epsilon)$: the ϵ-insensitive loss function.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the input space according to D, with labels $(c(\omega_1), \ldots, c(\omega_m))$.

- Problem : find a hypothesis $h_S : \mathscr{I} \to \mathscr{Y}$ in $\mathcal{H}$, i.e., a weight vector $\mathbf{w} \in \mathbb{R}^N$ and an offset $b \in \mathbb{R}$, which minimizes the following object function

$$F(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \left|(\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i)\right|_{\epsilon}.$$

Remarks

- The use of the ϵ-insensitive loss function leads to sparse solutions with a relatively small number of support vectors.
- The parameter C is a positive parameter determining the trade-off between the regularization term $\|\mathbf{w}\|^2$ and the empirical ϵ-insensitive loss.
- The objective function of the support vector regression differs from that of SVM only by the loss function.

- Using slack variables $\eta_i, \eta'_i, i \in [1, m]$, the ith empirical ϵ-insensitive loss becomes

$$
\begin{aligned}
& \left|(\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i)\right|_\epsilon \\
= \quad & \max(0, |(\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i)| - \epsilon) \\
= \quad & \min_{\eta_i, \eta'_i} (\eta_i + \eta'_i)
\end{aligned}
$$

subject to

$$
\begin{aligned}
& (\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) - \epsilon \leq \eta_i, \\
& c(\omega_i) - (\mathbf{w} \cdot \Phi(\omega_i) + b) - \epsilon \leq \eta'_i, \\
& \eta_i \geq 0, \\
& \eta'_i \geq 0.
\end{aligned}
$$

The Primal Problem of SVR

$$
\begin{aligned}
\text{Minimize} \quad & F(\mathbf{w}, b, \eta, \eta') = \tfrac{1}{2}\|\mathbf{w}\|^2 + C \textstyle\sum_{i=1}^{m} (\eta_i + \eta'_i) \\
\text{Subject to} \quad & (\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) - \epsilon - \eta_i \leq 0, \;\; i \in [1, m] \\
& c(\omega_i) - (\mathbf{w} \cdot \Phi(\omega_i) + b) - \epsilon - \eta'_i \leq 0, \;\; i \in [1, m] \\
& -\eta_i \leq 0, \;\; i \in [1, m] \\
& -\eta'_i \leq 0, \;\; i \in [1, m] \\
& \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \eta, \eta' \in \mathbb{R}^m.
\end{aligned}
$$

Qualification of the Primal Problem

- The object function $F(\mathbf{w}, b, \eta, \eta') = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\eta_i + \eta'_i)$ is infinitely differentiable and convex so that it is pseudoconvex at any feasible point.
- The $4m$ inequality constraint functions
  $g_i(\mathbf{w}, b, \eta, \eta') = (\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) - \epsilon - \eta_i$,
  $g'_i(\mathbf{w}, b, \eta, \eta') = c(\omega_i) - (\mathbf{w} \cdot \Phi(\omega_i) + b) - \epsilon - \eta'_i$,
  $h_i(\mathbf{w}, b, \eta, \eta') = -\eta_i$, $h'_i(\mathbf{w}, b, \eta, \eta') = -\eta'_i$, $i \in [1, m]$, are infinitely differentiable and linear so that it is pseudoconvex at any feasible point.

- $$\nabla F = \begin{bmatrix} \mathbf{w} \\ 0 \\ C\mathbf{1} \\ C\mathbf{1} \end{bmatrix},\ \nabla g_i = \begin{bmatrix} \Phi(\omega_i) \\ 1 \\ -\mathbf{e}_i \\ \mathbf{0} \end{bmatrix},\ \nabla g_i' = \begin{bmatrix} -\Phi(\omega_i) \\ -1 \\ \mathbf{0} \\ -\mathbf{e}_i \end{bmatrix},$$

  $$\nabla h_i = \begin{bmatrix} \mathbf{0} \\ 0 \\ -\mathbf{e}_i \\ \mathbf{0} \end{bmatrix} \text{ and } \nabla h_i' = \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \\ -\mathbf{e}_i \end{bmatrix},\ i \in [1, m].$$

- The Kuhn-Tucker necessary conditions are:

$$\nabla F + \sum_{i=1}^{m}(\lambda_i \nabla g_i + \lambda_i' \nabla g_i' + \mu_i \nabla h_i + \mu_i' \nabla h_i') = \mathbf{0}$$
$$\Leftrightarrow \mathbf{w} = \sum_{i=1}^{m}(\lambda_i' - \lambda_i)\Phi(\omega_i), \sum_{i=1}^{m}(\lambda_i' - \lambda_i) = 0,$$
$$C = \lambda_i + \mu_i, C = \lambda_i' + \mu_i', i \in [1, m];$$
$$\lambda_i g_i(\mathbf{w}, b, \eta, \eta') = 0, \lambda_i' g_i'(\mathbf{w}, b, \eta, \eta') = 0,$$
$$\mu_i \eta_i = 0, \mu_i' \eta_i' = 0, i \in [1, m];$$
$$\lambda_i, \lambda_i', \mu_i, \mu_i' \geq 0, \; i \in [1, m].$$

- Any feasible point $(\mathbf{w}, b, \eta, \eta')$ which satisfies the Kuhn-Tucker necessary conditions in above is a global minimum solution.
- The weight vector solution $\mathbf{w}^{SVR}$ is a linear combination of the training feature vectors $\Phi(\omega_1), \ldots, \Phi(\omega_m)$,

$$\mathbf{w}^{SVR} = \sum_{i=1}^{m}(\lambda_i'^{SVR} - \lambda_i^{SVR})\Phi(\omega_i).$$

Support Vectors

- Support vectors: any feature vector $\Phi(\omega_i)$ which appears in the linear combination $\mathbf{w}^{SVR} = \sum_{i=1}^{m} ({\lambda'_i}^{SVR} - \lambda_i^{SVR})\Phi(\omega_i)$, i.e., ${\lambda'_i}^{SVR} - \lambda_i^{SVR} \neq 0$.
- For each $i \in [1, m]$, at most one of ${\lambda'_i}^{SVR}$ and λ_i^{SVR} is nonzero, otherwise by the complementary slackness conditions, we have

  $$g_i(\mathbf{w}^{SVR}, b^{SVR}, \eta^{SVR}, {\eta'}^{SVR}) = 0 = g'_i(\mathbf{w}^{SVR}, b^{SVR}, \eta^{SVR}, {\eta'}^{SVR})$$

  so that $2\epsilon + \eta_i^{SVR} + {\eta'_i}^{SVR} = 0$, which is a contradiction.
- If $\lambda_i^{SVR} > 0$, the support vector $\Phi(\omega_i)$ satisfies

  $$(\mathbf{w}^{SVR} \cdot \Phi(\omega_i) + b^{SVR}) - c(\omega_i) = \epsilon + \eta_i^{SVR} \geq \epsilon$$

  and lies on or outside the ϵ-tube.
  - If $\eta_i^{SVR} = 0$, the support vector $\Phi(\omega_i)$ lies on the ϵ-tube,

i.e., $(\mathbf{w}^{SVR} \cdot \mathbf{x} + b^{SVR}) - \epsilon = c(\mathbf{x}_i)$.

  - If $\eta_i^{SVR} > 0$, the support vector $\Phi(\omega_i)$ lies outside the ϵ-tube and then by the complementary slackness conditions, $\mu_i^{SVR} = 0$ and then $\lambda_i^{SVR} = C$.

- If $\lambda_i'^{SVR} > 0$, the support vector $\Phi(\omega_i)$ satisfies

$$(\mathbf{w}^{SVR} \cdot \Phi(\omega_i) + b^{SVR}) - c(\omega_i) = -\epsilon - \eta_i'^{SVR} \leq -\epsilon$$

and lies on or outside the ϵ-tube.

  - If $\eta_i'^{SVR} = 0$, the support vector $\Phi(\omega_i)$ lies on the ϵ-tube, i.e., $(\mathbf{w}^{SVR} \cdot \mathbf{x} + b^{SVR}) + \epsilon = c(\mathbf{x}_i)$.
  - If $\eta_i'^{SVR} > 0$, the support vector $\Phi(\omega_i)$ lies outside the ϵ-tube and then by the complementary slackness conditions, $\mu_i'^{SVR} = 0$ and then $\lambda_i'^{SVR} = C$.

Remarks

- Support vectors fully define the SVR solution.
- Support vectors $\Phi(\omega_i)$ are either outliers, in which case either $\lambda_i^{SVR} = C$ or $\lambda_i'^{SVR} = C$, or vectors lying on the ϵ-tube.
- Feature vectors $\Phi(\omega_i)$ which are inside the ϵ-tube do not affect the solution to the SVR problem.
- When the number of feature vectors inside the ϵ-tube is relatively large, the hypothesis returned by SVR is a relatively sparse linear combination of feature vectors $\Phi(\omega_i)$.
- The choice of the parameter ϵ determines a trade-off between sparsity and accuracy: larger ϵ values provide sparser solutions, since more feature vectors can fall within the ϵ-tube, but may ignore too many key feature vectors for determining an accurate solution.

- While the solution $\mathbf{w}^{SVR}$ of the SVR problem is usually unique, the support vectors are not.

Determination of the Offset b^{SVR}

- For any $\lambda_j^{SVR} > 0$, i.e., $\Phi(\omega_j)$ being a support vector, we have

$$\begin{aligned} & b^{SVR} \\ = \quad & -\mathbf{w}^{SVR} \cdot \Phi(\omega_j) + c(\omega_j) + \epsilon + \eta_j^{SVR} \\ = \quad & -\sum_{i=1}^{m} (\lambda_i'^{SVR} - \lambda_i^{SVR})(\Phi(\omega_i) \cdot \Phi(\omega_j)) + c(\omega_j) + \epsilon + \eta_j^{SVR}. \end{aligned}$$

  - If $0 < \lambda_j^{SVR} < C$, then $\mu_j^{SVR} > 0$ and then $\eta_j^{SVR} = 0$ so that

$$b^{SVR} = -\sum_{i=1}^{m} (\lambda_i'^{SVR} - \lambda_i^{SVR})(\Phi(\omega_i) \cdot \Phi(\omega_j)) + c(\omega_j) + \epsilon.$$

- For any ${\lambda'_j}^{SVR} > 0$, i.e., $\Phi(\omega_j)$ being a support vector, we have

$$
\begin{aligned}
& b^{SVR} \\
= \quad & -\mathbf{w}^{SVR} \cdot \Phi(\omega_j) + c(\omega_j) - \epsilon - {\eta'_j}^{SVR} \\
= \quad & -\sum_{i=1}^{m} ({\lambda'_i}^{SVR} - \lambda_i^{SVR})(\Phi(\omega_i) \cdot \Phi(\omega_j)) + c(\omega_j) - \epsilon - {\eta'_j}^{SVR}.
\end{aligned}
$$

  - If $0 < {\lambda'_j}^{SVR} < C$, then ${\mu'_j}^{SVR} > 0$ and then ${\eta'_j}^{SVR} = 0$ so that

$$
b^{SVR} = -\sum_{i=1}^{m} ({\lambda'_i}^{SVR} - \lambda_i^{SVR})(\Phi(\omega_i) \cdot \Phi(\omega_j)) + c(\omega_j) - \epsilon.
$$

The Returned Hypothesis h_S^{SVR} by SVR

The returned hypothesis h_S^{SVR} by SVR is

$$
\begin{aligned}
h_S^{SVR}(\omega) &= \mathbf{w}^{SVR} \cdot \Phi(\omega) + b^{SVR} \\
&= \sum_{i=1}^{m} (\lambda_i'^{SVR} - \lambda_i^{SVR}) \Phi(\omega_i) \cdot \Phi(\omega) + b^{SVR} \\
&= \sum_{i=1}^{m} (\lambda_i'^{SVR} - \lambda_i^{SVR}) K(\omega_i, \omega) + b^{SVR},
\end{aligned}
$$

where

$$K(\omega_i, \omega) \triangleq \Phi(\omega_i) \cdot \Phi(\omega)$$

is the PDS kernel associated with the feature mapping Φ.

- When $0 < \lambda_j^{SVR} < C$,
  $b^{SVR} = -\sum_{i=1}^{m} (\lambda_i'^{SVR} - \lambda_i^{SVR}) K(\omega_i, \omega_j) + c(\omega_j) + \epsilon$.

- When $0 < {\lambda'}_j^{SVR} < C$,
  $b^{SVR} = -\sum_{i=1}^{m}({\lambda'}_i^{SVR} - \lambda_i^{SVR})K(\omega_i, \omega_j) + c(\omega_j) - \epsilon.$
- With this formulation, SVR can be extended to an arbitrary PDS kernel K over the input space $\mathscr{I}$, where a Hilbert space $\mathbb{H}$ and a feature mapping $\Phi : \mathscr{I} \to \mathbb{H}$ can be associated.
- We will use the Lagrangian dual problem to determine optimal $\lambda_i^{SVR}, {\lambda'}_i^{SVR}$.

Lagrangian Dual Function for SVR

- $X = \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^m$: a nonempty open convex set.
- Lagrangian function: for all $\mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \eta, \eta' \in \mathbb{R}^m$ and $\lambda, \lambda', \mu, \mu' \in \mathbb{R}^m$,

$$
\begin{aligned}
& L(\mathbf{w}, b, \eta, \eta', \lambda, \lambda', \mu, \mu') \\
= \quad & F(\mathbf{w}, b, \eta, \eta') + \sum_{i=1}^{m} (\lambda_i g_i(\mathbf{w}, b, \eta, \eta') + \lambda_i' g_i'(\mathbf{w}, b, \eta, \eta') + \\
& \mu_i h_i(\mathbf{w}, b, \eta, \eta') + \mu_i' h_i'(\mathbf{w}, b, \eta, \eta')) \\
= \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{m} (\eta_i + \eta_i') + \\
& \sum_{i=1}^{m} (\lambda_i ((\mathbf{w} \cdot \Phi(\omega_i) + b) - c(\omega_i) - \epsilon - \eta_i) + \\
& \lambda_i' (c(\omega_i) - (\mathbf{w} \cdot \Phi(\omega_i) + b) - \epsilon - \eta_i') - \mu_i \eta_i - \mu_i' \eta_i').
\end{aligned}
$$

- For any fixed $\lambda, \lambda', \mu, \mu' \in \mathbb{R}^m$, the gradient ∇L of the Lagrangian function w.r.t. $(\mathbf{w}, b, \eta, \eta')$ is

$$
\begin{aligned}
\nabla L &= \nabla F + \sum_{i=1}^{m} (\lambda_i \nabla g_i + \lambda_i' \nabla g_i' + \mu_i \nabla h_i + \mu_i' \nabla h_i') \\
&= \begin{bmatrix} \mathbf{w} \\ 0 \\ C\mathbf{1} \\ C\mathbf{1} \end{bmatrix} + \sum_{i=1}^{m} \left(\lambda_i \begin{bmatrix} \Phi(\omega_i) \\ 1 \\ -\mathbf{e}_i \\ \mathbf{0} \end{bmatrix} + \lambda_i' \begin{bmatrix} -\Phi(\omega_i) \\ -1 \\ \mathbf{0} \\ -\mathbf{e}_i \end{bmatrix} + \right. \\
& \qquad \left. \mu_i \begin{bmatrix} \mathbf{0} \\ 0 \\ -\mathbf{e}_i \\ \mathbf{0} \end{bmatrix} + \mu_i' \begin{bmatrix} \mathbf{0} \\ 0 \\ \mathbf{0} \\ -\mathbf{e}_i \end{bmatrix} \right)
\end{aligned}
$$

and the Hessian matrix is

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_{N\times N} & \mathbf{0}_{N\times(1+2m)} \\ \mathbf{0}_{(1+2m)\times N} & \mathbf{0}_{(1+2m)\times(1+2m)} \end{bmatrix}$$

which is positive semi-definite.

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

$$\mathbf{w} = \sum_{i=1}^{m}(\lambda_i' - \lambda_i)\Phi(\omega_i), \sum_{i=1}^{m}(\lambda_i' - \lambda_i) = 0,$$
$$C = \lambda_i + \mu_i, C = \lambda_i' + \mu_i', i \in [1, m].$$

  – Note that for any fixed $\lambda, \lambda', \mu, \mu' \in \mathbb{R}^m$, $\sum_{i=1}^{m}(\lambda_i' - \lambda_i) \neq 0$ or $C \neq \lambda_i + \mu_i$ for some $i \in [1, m]$ or $C \neq \lambda_i' + \mu_i'$ for some

$i \in [1, m]$ if and only if the infimum of the Lagrangian function $L(\mathbf{w}, b, \eta, \eta', \lambda, \lambda', \mu, \mu')$ is $-\infty$.

- Lagrangian dual function: for any $\lambda, \lambda', \mu, \mu' \in \mathbb{R}^m$,

$$
\begin{aligned}
& \theta(\lambda, \lambda', \mu, \mu') \\
= \quad & \inf_{(\mathbf{w}, b, \eta, \eta') \in X} L(\mathbf{w}, b, \eta, \eta', \lambda, \lambda', \mu, \mu') \\
= \quad & \begin{cases}
\frac{1}{2}\|\hat{\mathbf{w}}\|^2 + C \sum_{i=1}^m (\hat{\eta}_i + \hat{\eta}'_i) + \\
\sum_{i=1}^m (\lambda_i((\hat{\mathbf{w}} \cdot \Phi(\omega_i) + \hat{b}) - c(\omega_i) - \epsilon - \hat{\eta}_i) + \\
\lambda'_i(c(\omega_i) - (\hat{\mathbf{w}} \cdot \Phi(\omega_i) + \hat{b}) - \epsilon - \hat{\eta}'_i) - \mu_i \hat{\eta}_i - \mu'_i \hat{\eta}'_i), \\
\text{ if } \sum_{i=1}^m (\lambda'_i - \lambda_i) = 0, C = \lambda_i + \mu_i, C = \lambda'_i + \mu'_i, i \in [1, m], \\
-\infty, \text{ otherwise}
\end{cases} \\
= \quad & \begin{cases}
-\epsilon \sum_{i=1}^m (\lambda'_i + \lambda_i) + \sum_{i=1}^m (\lambda'_i - \lambda_i) c(\omega_i) \\
-\frac{1}{2} \sum_{i,j=1}^m (\lambda'_i - \lambda_i)(\lambda'_j - \lambda_j)(\Phi(\omega_i) \cdot \Phi(\omega_j)), \\
\text{ if } \sum_{i=1}^m (\lambda'_i - \lambda_i) = 0, C = \lambda_i + \mu_i, C = \lambda'_i + \mu'_i, i \in [1, m], \\
-\infty, \text{ otherwise.}
\end{cases}
\end{aligned}
$$

Lagrangian Dual Problem for SVR

Maximize $\theta(\lambda, \lambda', \mu, \mu') = -\epsilon \sum_{i=1}^{m}(\lambda_i' + \lambda_i) + \sum_{i=1}^{m}(\lambda_i' - \lambda_i)c(\omega_i)$

$-\frac{1}{2}\sum_{i,j=1}^{m}(\lambda_i' - \lambda_i)(\lambda_j' - \lambda_j)(\Phi(\omega_i) \cdot \Phi(\omega_j)),$

Subject to $\lambda_i, \lambda_i', \mu_i, \mu_i' \geq 0, i \in [1, m]$

$\lambda_i + \mu_i - C = 0, i \in [1, m]$

$\lambda_i' + \mu_i' - C = 0, i \in [1, m]$

$\sum_{i=1}^{m}(\lambda_i' - \lambda_i) = 0$

$\lambda, \lambda', \mu, \mu' \in \mathbb{R}^m$

Or equivalently,

$$
\begin{aligned}
\text{Maximize} \quad & \theta(\lambda, \lambda') = -\epsilon(\lambda' + \lambda)^T \mathbf{1} + (\lambda' - \lambda)^T \mathbf{y} \\
& -(\lambda' - \lambda)^T \mathbf{K} (\lambda' - \lambda), \\
\text{Subject to} \quad & \lambda_i, \lambda'_i \geq 0, i \in [1, m] \\
& C - \lambda_i \geq 0, i \in [1, m] \\
& C - \lambda'_i \geq 0, i \in [1, m] \\
& \textstyle\sum_{i=1}^{m} (\lambda'_i - \lambda_i) = 0 \\
& \lambda, \lambda' \in \mathbb{R}^m,
\end{aligned}
$$

where $\mathbf{y} = [c(\omega_1), \ldots, c(\omega_m)]^T$ is the label vector and $\mathbf{K} = [K(\omega_i, \omega_j)]$ with $K(\omega_i, \omega_j) = \Phi(\omega_i) \cdot \Phi(\omega_j)$ is the kernel matrix associated with the sample $S = (\omega_1, \ldots, \omega_m)$.

- This dual problem can be solved by the sequential minimal optimization (SMO) algorithm.

The Contents of This Lecture

- Generalization bounds
- Linear regression
- Kernel ridge regression
- Support vector regression
- Lasso

Least Absolute Shrinkage and Selection Operator (Lasso) Problem

- c: a fixed but unknown target concept in a concept class $\mathcal{C}$ of real-valued measurable function on the input space $\mathscr{I}$.
- $\Phi : \mathscr{I} \to \mathbb{R}^N$: a feature mapping from the input space $\mathscr{I}$ to the N-dimensional feature space $\mathbb{R}^N$.
- $\mathcal{H} = \{\omega \mapsto \mathbf{w} \cdot \Phi(\omega) + b \mid \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}\}$: the hypothesis set of all affine functions in the feature space $\mathbb{R}^N$.
- $L(y', y) = (y' - y)^2$: the squared-error loss function.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the input space according to D, with labels $(c(\omega_1), \ldots, c(\omega_m))$.

- Problem : find a hypothesis $h_S : \mathscr{I} \to \mathscr{Y}$ in $\mathcal{H}$, i.e., a weight vector $\mathbf{w} \in \mathbb{R}^N$ and an offset $b \in \mathbb{R}$, which minimizes the following object function

$$F(\mathbf{w}, b) = \lambda \|\mathbf{w}\|_1 + \sum_{i=1}^{m} \left(\mathbf{w} \cdot \Phi(\omega_i) + b - c(\omega_i)\right)^2 .$$

  – This is a convex optimization problem.

Remarks

- The parameter λ is a positive parameter determining the trade-off between the regularization term $\|\mathbf{w}\|_1$ and the empirical mean squared error.

- Except for the shift $b = 0$ in the second term, the objective function of the Lasso problem differs from that of kernel ridge regression only by the first term, where L_1-norm is used instead of the square of the L_2-norm.

- Unlike the KRR and SVR algorithms, the Lasso algorithm does not admit a natural use of PDS kernels.

Alternative Formulations of Lasso Problem

$$
\begin{array}{ll}
\text{Minimize} & F(\mathbf{w}, b) = \sum_{i=1}^{m} \left(\mathbf{w} \cdot \Phi(\omega_i) + b - c(\omega_i)\right)^2 \\
\text{Subject to} & \|\mathbf{w}\|_1 - \Lambda_1 \leq 0 \\
& \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}.
\end{array}
$$

or equivalently, using slack variables $\eta_i = c(\omega_i) - \mathbf{w} \cdot \Phi(\omega_i) - b$, $i \in [1, m]$,

$$
\begin{array}{ll}
\text{Minimize} & F(\mathbf{w}, b, \eta) = \frac{1}{2} \sum_{i=1}^{m} \eta_i^2 \\
\text{Subject to} & \sum_{j=1}^{N} (-1)^{k_j} w_j - \Lambda_1 \leq 0, k_1, \ldots, k_N \in [0, 1] \\
& c(\omega_i) - \mathbf{w} \cdot \Phi(\omega_i) - b - \eta_i = 0, \ i \in [1, m] \\
& \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}, \eta \in \mathbb{R}^m.
\end{array}
$$

Main Idea of Lasso

- The key property of Lasso, as in the case of other algorithms using the L_1 norm constraint, is that it leads to a sparse solution $\mathbf{w}^{Lasso}$, that is one with few non-zero components as shown in Figure 10.6 of the textbook.

Converting Lasso Problem to a Smaller QP Problem

- Any real number w can be written as the difference of two non-negative numbers w^+, w^-, i.e., $w = w_i^+ - w_i^-$. There infinitely many such pairs (w^+, w^-) for w and

$$|w| = \min_{w^+, w^- \geq 0, w = w^+ - w^-} (w^+ + w^-).$$

Now the minimization problem of Lasso becomes

$$\begin{aligned}
& \min_{\mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}} F(\mathbf{w}, b) \\
= \quad & \min_{\mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R}} \lambda \sum_{j=1}^{N} |w_j| + \sum_{i=1}^{m} (\mathbf{w} \cdot \Phi(\omega_i) + b - c(\omega_i))^2 \\
= \quad & \min_{\mathbf{w}^+, \mathbf{w}^- \geq \mathbf{0}, b \in \mathbb{R}} \lambda (\mathbf{w}^+ + \mathbf{w}^-)^T \mathbf{1} + \sum_{i=1}^{m} \big((\mathbf{w}^+ - \mathbf{w}^-) \cdot \Phi(\omega_i) + b - c(\omega_i)\big)^2.
\end{aligned}$$