Mathematical Theory of Neural Network Models

Lecture 1: An Introduction to Supervised Learning

July 2

Lecturer: Lei Wu *Scribe: Shuhai Zhao, Yilei Han*

1 Supervised Learning

Some basic terminologies:

- *features*: The set of attibutes, often represented as a vector, associated to an example.
- *Hypothesis space*: A set $\mathcal{F}$ of functions mapping features to the set of labels $\mathcal{Y}$.
- *Loss function*: A function l that measures the difference, or loss, between a predicted label and a true label: $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, for example, $l(y, y') = (y - y')^2$.

Problems in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as *supervised learning* problems. Specifically, in the most general scenario of supervised learning, the distribution D is defined over $\mathcal{X} \times \mathcal{Y}$, and the training data is a labeled sample S drawn i.i.d. according to D:

$$S = ((\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)),$$

where $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}, i = 1, 2, ..., n$, $y_i = f^*(\boldsymbol{x}_i) + \epsilon_i$.

Definition 1:
Given a hypothesis f, a loss function l, and a sample $S = \{(x_1, y_1), ..., (x_n, y_n)\}$, the *empirical risk* (ER) is defined by

$$L_n(f) = \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i),$$

and the *population risk* is defined by

$$L(f) = \mathbb{E}_{(x,y)\sim D}(l(f(x), y)).$$

In supervised learning, we consider the *empirical risk minimization*(ERM) problem:

$$\hat{f} = \text{argmin}_{f\in\mathcal{F}}\, L_n(f).$$

If $\mathcal{F}$ is parameterized by $\boldsymbol{\theta}$, i.e., $\mathcal{F} = \{f(\boldsymbol{x}; \boldsymbol{\theta}) | \boldsymbol{\theta} \in \Lambda\}$, we can also write

$$\hat{\boldsymbol{\theta}}_n := \text{argmin}_{\boldsymbol{\theta}}\, L_n(\boldsymbol{\theta}),$$

and

$$\boldsymbol{\theta}^* := \text{argmin}_{\boldsymbol{\theta}}\ L(\boldsymbol{\theta}).$$

We define the *generalization error* to be

$$f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_n) - f^* = (f(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_n) - f(\boldsymbol{x}; \boldsymbol{\theta}^*)) + (f(\boldsymbol{x}; \boldsymbol{\theta}^*) - f^*).$$

On the right hand side of the equality, we call the first term *estimation error*, and the second term *approximation error*.

As the ERM problem is sometimes difficult to solve, so we also consider structure ERM:

$$\hat{\boldsymbol{\theta}}_n(\epsilon) = \text{argmin}_{L_n(\boldsymbol{\theta}) \leq \epsilon}\ \|\boldsymbol{\theta}\|$$

or

$$\hat{\boldsymbol{\theta}}_n(\lambda) = \text{argmin}_{\boldsymbol{\theta}}\ (L_n(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|^2).$$

In the following sections, we mainly concentrate on two questions:

- How to choose our hypothesis space?
- how to learn from the chosen hypothesis space?

2 Linear Regression

Let $\mathcal{F} = \{\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x} | \boldsymbol{\beta} \in \mathbb{R}^d\}$, $l(y, y') = (y - y')^2$, the ERM problem becomes

$$\min \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i - y_i)^2,$$

let $Z = (\boldsymbol{x}_1^{\mathrm{T}}, ..., \boldsymbol{x}_n^{\mathrm{T}})^{\mathrm{T}}$, the problem can be written as

$$\min \frac{1}{n} \|Z\boldsymbol{\beta} - Y\|^2$$

• For $n < d$, we consider *Ridge regression*:

$$\min \frac{1}{n} \|Z\boldsymbol{\beta} - Y\|^2 + \lambda \|\boldsymbol{\beta}\|^2$$

3 Generalized Linear Model

Let

$$\mathcal{F} = \{ \sum_{i=1}^{m} \beta_j \phi_j(x) | \boldsymbol{\beta} \in \mathbb{R}^m \},$$

where the ϕ_j's are basis functions(for example, polynomials, splines, wavelets,...). Write $\Phi = (\phi_j(x_i))_{i,j}$, the ERM problem becomes

$$\hat{\boldsymbol{\beta}} = \text{argmin } \|\Phi\boldsymbol{\beta} - Y\|^2 + \lambda\|\boldsymbol{\beta}\|^2,$$

by differentiating with respect to $\boldsymbol{\beta}$, we get

$$\hat{\boldsymbol{\beta}} = (\Phi^{\mathrm{T}}\Phi + \lambda I)^{-1}\Phi^{\mathrm{T}}y.$$

By easy calculations:

$$\hat{\boldsymbol{\beta}} = \Phi^{\mathrm{T}}(\Phi\Phi^{\mathrm{T}} + \lambda I)^{-1}y.$$

Denote $\alpha := (\Phi\Phi^{\mathrm{T}} + \lambda I)^{-1}y$ and consider the case: $m \gg n$
The complexity of calculating the inverse of a *m*-order matrix is O(m^3) or O(Cm^2). By the above transform, the order of the matrix which will be inversed is decreased to n and the complexity is decreased to O(n^3).

The key observations:

- $\hat{\boldsymbol{\beta}} \in span\{\boldsymbol{\phi}(\boldsymbol{x}_1), ..., \boldsymbol{\phi}(\boldsymbol{x}_n)\}$, we solve

$$\min \|\Phi\Phi^{\mathrm{T}}\alpha - Y\|^2 + \lambda\|\Phi^{\mathrm{T}}\alpha\|^2,$$

- The learned model is given by

$$f(\boldsymbol{x}; \hat{\boldsymbol{\beta}}) = \Phi(\boldsymbol{x})\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \alpha_i \langle \boldsymbol{\phi}(\boldsymbol{x}_i), \boldsymbol{\phi}(\boldsymbol{x})\rangle$$

  Let $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}')\rangle$, we have

$$f(\boldsymbol{x}; \hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}), \tag{1}$$

  with the coefficients are given by

$$\boldsymbol{\alpha} = (K + \lambda I_n)^{-1}Y, \tag{2}$$

  where $K = (k(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j} \in \mathbb{R}^{n\times n}$. This matrix is called Gram or Kernel matrix.

4 Kernel Method

After the observations in , we have A kernel over $\mathcal{X}$ is a function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which satisfies two conditions:

- For any $x, x' \in \mathcal{X}$, $k(x, x') = k(x', x)$;
- For any $x_1, x_2, ..., x_n \in \mathcal{X}$, the Gram matrix $K = (k(x_i, x_j))$ is positive semidefinite.

Kernel Ridge Regression: For any kernel k, the estimator (1) is called kernel ridge regression.

Example:

- $k(\boldsymbol{x}, \boldsymbol{x}') = k_0(\boldsymbol{x} - \boldsymbol{x}'), k(\boldsymbol{x}, \boldsymbol{x}') = k_0(\|\boldsymbol{x} - \boldsymbol{x}'\|_2)$.
- $k(u, v) = (\langle u, v\rangle + 1)^m$ Denote $u_0 = v_0 = 1$, then

$$k(u, v) = \sum_{J\in\{0,1,\dots,d\}^m} (\prod_{i=1}^{m} u_{J_i})(\prod_{i=1}^{m} v_{J_i}) = \langle \phi(u), \phi(v)\rangle$$

- $k(x, x') = e^{-\frac{(x-x')^2}{2}}$ From the Taylor expansion, we have

$$k(x, x') = \sum_{n=0}^{\infty} \frac{1}{\sqrt{n!}} e^{-\frac{x^2}{2}} x^n \frac{1}{\sqrt{n!}} e^{-\frac{x'^2}{2}} x'^n = \sum_{n=0}^{\infty} \phi_n(x)\phi_n(x')$$

5 Random Fourier Feature model

Shift invariant kernel is equivalent to *Random Fourier feature*. So for a high dimensional problem, it's effective to use the latter to decrease the dimension and therefore *Kernel Method* could be used. The argument is promised by the below theorem.

Theorem 5.1 (Bochner)**.** *Assume $k(\boldsymbol{x}, \boldsymbol{x}') = k_0(\boldsymbol{x} - \boldsymbol{x}')$. Then k is a continuous kernel function iff k_0 is the Fourier transform of a non-negative measure.*

The probability measure is an example:
$k_0(\boldsymbol{x} - \boldsymbol{y}) = \int_{\mathbb{R}^d} p(\boldsymbol{\omega}) e^{\mathrm{i}\boldsymbol{\omega}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y})} \mathrm{d}\boldsymbol{\omega} = \mathbb{E}_{\boldsymbol{\omega}\sim\mathbb{P}}[z_{\boldsymbol{\omega}}(\boldsymbol{x})\bar{z}_{\boldsymbol{\omega}}(\boldsymbol{x})]$. Where $z_{\boldsymbol{\omega}}(\boldsymbol{x}) = e^{\mathrm{i}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}}$
The vector form is $z_{\boldsymbol{\omega}}(\boldsymbol{x}) = \begin{pmatrix} \cos(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}) \\ \sin(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}) \end{pmatrix}$

An example for the theorem: $z_{\boldsymbol{\omega}}(\boldsymbol{x}) = \sqrt{2}\cos(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x} + b)$,$b \sim u([0, 2\pi])$

$$\begin{aligned} \frac{1}{2\pi}\int_0^{2\pi} z_{\boldsymbol{\omega}}(\boldsymbol{x}, b) z_{\boldsymbol{\omega}}(\boldsymbol{y}, b)\mathrm{d}b &= \frac{2}{2\pi}\int_0^{2\pi} \cos(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b)\cos(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{y} + b)\mathrm{d}b & (3)\\ &= \frac{1}{\pi}\int_0^{2\pi} \cos(\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x} + \boldsymbol{y}) + 2b) + \cos(\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}))\mathrm{d}b & (4)\\ &= 2\cos(\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y})) & (5) \end{aligned}$$

The random approximation of the kernel:

$$k(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{\omega}\sim\mathbb{P}}[z_{\boldsymbol{\omega}}(\boldsymbol{x}) z_{\boldsymbol{\omega}}(\boldsymbol{y})] \simeq \frac{1}{m}\sum_{k=1}^{m} z_{\boldsymbol{\omega}_k}(\boldsymbol{x}) z_{\boldsymbol{\omega}_k}(\boldsymbol{y})$$

The difference between *Emperical Risk* and *Population Risk* could be estimated by the Central limit theorem.

$\sup_{x,y\in M} |k(\boldsymbol{x},\boldsymbol{y}) - \frac{1}{m}\sum_{k=1}^{m} z_{\boldsymbol{\omega}_k}(\boldsymbol{x}) z_{\boldsymbol{\omega}_k}(\boldsymbol{y})| \sim \frac{C}{\sqrt{m}} = \epsilon, C(diam(M), d)$

One could refer to *Random Features for Large-Scale kernel Machines*(2007,Ali Rahimi,Benjamin Recht) for more details.