Lecture-18: Regression

1 Regression

Consider input space $\mathcal{X} \subseteq \mathbb{R}^d$ and output space $\mathcal{Y} \subseteq \mathbb{R}$, and an *i.i.d.* labeled sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ from an unknown distribution $\mathcal{D}$. The deterministic case when $S_{\mathcal{X}}$ is *i.i.d.* with an unknown distribution $\mathcal{D}_{\mathcal{X}}$ and $Y_i = c(X_i)$ for an unknown concept $c : \mathcal{X} \to \mathcal{Y}$, is a special case. We consider a loss function $L : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_+$ that measures the magnitude of error.

Example 1.1. A common loss function used in regression problems is $L_p : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_+$ for $p \geqslant 1$, defined by $L_p(y', y) \triangleq |y' - y|^p$ for all $(y', y) \in \mathcal{Y}' \times \mathcal{Y}$.

Definition 1.2. The hypothesis class is denoted by $\mathcal{H} \subseteq (\mathcal{Y}')^{\mathcal{X}}$. The generalization error for a hypothesis $h \in \mathcal{H}$ under loss function L is defined as $R(h) \triangleq \mathbb{E}_{\mathcal{D}} L(h(X), c(X))$. The empirical error for a hypothesis $h \in \mathcal{H}$ under loss function L, and sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ is defined as $\hat{R}(h) \triangleq \frac{1}{m} \sum_{i=1}^m L(h(x_i), y_i)$.

Assumption 1.3. We consider *bounded regression problems* where loss functions are bounded. That is, some $M > 0$, we have $L(h(x), c(x)) \leqslant M$ for all inputs $x \in \mathcal{X}$ and hypothesis $h \in \mathcal{H}$.

1.1 Generalization bounds

We are interested in generalization bounds on the generalization error.

1.1.1 Finite hypothesis set

Theorem 1.4 (Hoeffding). *Consider a random walk $S : \Omega \to \mathbb{R}^{\mathbb{N}}$ defined by $S_n \triangleq \sum_{i=1}^n X_i$ for each $n \in \mathbb{N}$, where the random step-size sequence $X : \Omega \to \prod_{i \in \mathbb{N}} [a_i, b_i]$ is independent and bounded. Then,*

$$P\{S_m - \mathbb{E}S_m \leqslant -\epsilon\} \leqslant e^{-\frac{2\epsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}}.$$

Theorem 1.5. *Consider an M-bounded loss function L and finite hypothesis set $\mathcal{H}$. For a fixed $\delta > 0$, with probability at least $1 - \delta$, we have $R(h) \leqslant \hat{R}(h) + M\sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}$, for any hypothesis $h \in \mathcal{H}$.*

Proof. Applying the Hoeffding inequality for M-bounded loss functions, we can bound the probability of generalization error exceeding empirical error by at least ϵ, as

$$P\{R(h) - \hat{R}(h) \geqslant \epsilon\} = P\left\{\sum_{i=1}^m [L(h(X_i), Y_i) - \mathbb{E}L(h(X_i), c(X_i))] \leqslant -m\epsilon\right\} \leqslant e^{-\frac{2m\epsilon^2}{M^2}}.$$

It follows from the union bound that

$$P\Big(\cup_{h \in \mathcal{H}} \{R(h) - \hat{R}(h) \geqslant \epsilon\}\Big) \leqslant \sum_{h \in \mathcal{H}} P\{R(h) - \hat{R}(h) \geqslant \epsilon\} \leqslant |\mathcal{H}| e^{-\frac{2m\epsilon^2}{M^2}}.$$

Defining $\delta \triangleq |\mathcal{H}| e^{-\frac{2m\epsilon^2}{M^2}}$, the result follows. □

1.1.2 Infinite hypothesis set

Theorem 1.6. *Consider a family of functions* $\mathcal{G} \triangleq \{(x,y) \mapsto L(h(x),y) : h \in \mathcal{H}\}$. *For any* $\delta > 0$, *with probability at least* $1-\delta$, *we have* $R(h) \leqslant \hat{R}(h) + 2\hat{\mathfrak{R}}_m(\mathcal{G}) + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$, *for any hypothesis* $h \in \mathcal{H}$.

Proof. We observe that $\mathbb{E}g(z) = R(h) = \mathbb{E}_{\mathcal{D}} L(h(X),c(X))$ and $\frac{1}{m}\sum_{i=1}^m g(z_i) = \hat{R}(h)$. □

Remark 1. For $p \geqslant 1$, we write the loss function $L_p(y',y) \triangleq |y'-y|^p$ for all $(y',y) \in \mathcal{Y}' \times \mathcal{Y}$. For the loss function to be bounded, we assume that $|h(x)-c(x)| \leqslant M$ for all $x \in \mathcal{X}$. For this loss function, we call $\mathcal{G} = \mathcal{H}_p$.

Remark 2. Recall Talagrand's contraction lemma that states that for a hypothesis set $\mathcal{H}$ and another set $\hat{\mathcal{H}} \triangleq \{\Phi \circ h : h \in \mathcal{H}\}$ where Φ is L-Lipschitz, we have $\hat{\mathfrak{R}}_S(\hat{\mathcal{H}}) \leqslant L\hat{\mathfrak{R}}_S(\mathcal{H})$.

Theorem 1.7. $\hat{\mathfrak{R}}_S(\mathcal{H}_p) \leqslant pM^{p-1}\hat{\mathfrak{R}}_S(\mathcal{H})$ *and* $\mathfrak{R}_m(\mathcal{H}_p) \leqslant pM^{p-1}\mathfrak{R}_m(\mathcal{H})$.

Proof. Defining $\mathcal{H}' \triangleq \{x \mapsto h(x) - c(x) : h \in \mathcal{H}\}$, we observe that $\mathcal{H}_p = \{\Phi_p \circ h : h \in \mathcal{H}'\}$, where $\Phi_p : \mathbb{R} \to \mathbb{R}_+$ defined by $\Phi_p(x) = |x|^p$. We observe that Φ_p is pM^{p-1}-Lipschitz for all $x \in [0,M]$. Therefore, it follows from Talagrand's Lemma, that $\hat{\mathfrak{R}}_S(\mathcal{H}_p) \leqslant pM^{p-1}\hat{\mathfrak{R}}_S(\mathcal{H}')$, where

$$\hat{\mathfrak{R}}_S(\mathcal{H}') = \frac{1}{m}\mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i(h(x_i)-c(x_i))\Big] = \frac{1}{m}\mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \sigma_i(h(x_i)\Big] = \hat{\mathfrak{R}}_S(\mathcal{H}).$$

□

Remark 3. Let $p \geqslant 1$ and $|h(x)-c(x)| \leqslant M$ for all $x \in \mathcal{X}$ and $h \in \mathcal{H}$. Then, for any $y' \in \mathcal{Y}'$, the map $y' \mapsto |y'-y|^p$ is pM^{p-1}-Lipschitz for $(y'-y) \in [-M,M]$. Therefore, for a fixed $\delta > 0$, with probability at least $1-\delta$,

$$\mathbb{E}\,|h(X)-Y|^p \leqslant \frac{1}{m}\sum_{i=1}^m |h(x_i)-y_i|^p + 2pM^{p-1}\hat{\mathfrak{R}}_S(\mathcal{H}) + M^p\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

2 Regression algorithms

2.1 Kernel ridge regression

We consider regression for bounded linear hypotheses in a feature space $\mathbb{H}$ defined by a feature map $\Phi : \mathcal{X} \to \mathbb{H}$ associate to a PDS kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that $\langle \Phi(x), \Phi(x')\rangle = K(x,x')$ for all $x,x' \in \mathcal{X}$.

Theorem 2.1. *Let* $\mathcal{H} \triangleq \{x \mapsto \langle w, \Phi(x)\rangle : \|w\|_{\mathbb{H}} \leqslant \Lambda\}$, *PDS kernel* $K(x,x) \leqslant r^2$ *for all* $x \in \mathcal{X}$ *and some* $r > 0$, *and* $|h(x)-y| < M$ *for all* $(x,y) \in \mathcal{X} \times \mathcal{Y}$ *some* $M > 0$. *Then, for a fixed* $\delta > 0$, *with probability at least* $1-\delta$,

$$R(h) \leqslant \hat{R}(h) + 4M\sqrt{\frac{r^2\Lambda^2}{m}} + M^2\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \text{ for all } h \in \mathcal{H}.$$

Remark 4. The learning bound suggests minimizing a trade-off between the empirical squared loss, and the norm of the weight vector or equivalently the norm squared.

Definition 2.2 (Kernel ridge regression). The optimal weight vector is defined as

$$\arg\min_{\mathbf{w}} F(\mathbf{w}) = \arg\min_{\mathbf{w}} \lambda\,\|\mathbf{w}\|^2 + \sum_{i=1}^m (\langle \mathbf{w}, \Phi(x_i)\rangle - y_i)^2.$$

Remark 5. The Kernel ridge regression problem is a convex optimization problem with quadratic cost, and the optimal solution can be obtained by taking derivatives with respect to weight vector $\mathbf{w}$, to obtain

$$0 = \nabla_{\mathbf{w}} F(\mathbf{w}) = 2\lambda\mathbf{w} + \sum_{i=1}^m 2(\langle \mathbf{w}, \Phi(x_i)\rangle - y_i)\Phi(x_i).$$

For N-dimensional feature space $\mathbb{H}$, we can define matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{N\times m}$ by $\mathbf{X} \triangleq \begin{bmatrix}\Phi(x_1) & \dots & \Phi(x_m)\end{bmatrix}$ and $\mathbf{Y} \triangleq \begin{bmatrix}y_1 & \dots & y_m\end{bmatrix}$, to rewrite the solution to Kernel ridge regression as

$$0 = 2\lambda\mathbf{w} + 2\mathbf{X}(\mathbf{X}^T\mathbf{w} - \mathbf{Y}).$$

Since $\mathbf{X}\mathbf{X}^T$ is positive semidefinite, it follows that $(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})$ is positive definite, and its inverse exists. Therefore, we have $\mathbf{w} = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{Y}$. Computing the optimal weight vector requires $O(mN^2 + N^3)$ computations.

Remark 6. The positive parameter λ determines the trade-off between the regularization term $\|\mathbf{w}\|^2$ and the empirical mean squared error. Kernel ridge regression can also be written as a constrained optimization problem

$$\arg\min_{\mathbf{w}} \left\{ \sum_{i=1}^{m} (\langle \mathbf{w}, \Phi(x_i) \rangle - y_i)^2 : \|\mathbf{w}\| \leqslant \Lambda \right\}.$$

Remark 7. Kernel ridge regression has (a) good theoretical guarantees, (b) good stability properties, (c) closed form optimal solution, and (c) can be generalized to maps $c : \mathcal{X} \to \mathbb{R}^p$ by formulating the problem as p independent regression problems. However, Kernel ridge regression is computationally expensive and optimal weight vector is not sparse in general.

2.2 Support vector regression

This algorithm is inspired by SVM, where the points that are ϵ-close to the predicted output are not penalized and points further away are penalized according to their distance from the predicted output.

Definition 2.3. Consider the ϵ-insensitive loss $|\cdot|_\epsilon : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_+$ defined by $|y' - y|_\epsilon = \max\{0, |y' - y| - \epsilon\}$ for all $y', y \in \mathcal{Y} \subseteq \mathbb{R}$.

Remark 8. The ϵ-insensitive loss function provides sparse solutions, where the insensitivity parameter ϵ controls the sparsity. Sparsity increases with the insensitivity ϵ.

Definition 2.4 (Support vector regression). For a hypothesis set of linear functions

$$\mathcal{H} \triangleq \left\{ x \mapsto \langle \mathbf{w}, \Phi(x) \rangle + b : \mathbf{w} \in \mathbb{R}^N, b \in \mathbb{R} \right\},$$

the optimal hypothesis is defined as

$$\arg\min_{\mathbf{w},b} F(\mathbf{w}) = \arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m} |y_i - (\langle \mathbf{w}, \Phi(x_i) \rangle + b)|_\epsilon.$$

Remark 9. Support vector regression is a convex quadratic optimization with affine constraints. It differs from Kernel ridge regression in that the loss functions is akin to L_1 and not L_2. One can choose any PDS K for this problem, however, one still needs to choose regularization parameter C and insensitivity parameter ϵ.

Theorem 2.5. *Let $\mathcal{H} \triangleq \{x \mapsto \langle \mathbf{w}, \Phi(x) \rangle : \|\mathbf{w}\|_{\mathbb{H}} \leqslant \Lambda\}$, kernel function $K(x,x) \leqslant r^2$ for all $x \in \mathcal{X}$, and $|h(x) - y| \leqslant M$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. For ϵ-insensitive loss function and some $\delta > 0$, with probability at least $1 - \delta$, we have*

$$R(h) \leqslant \hat{R}(h) + 2\sqrt{\frac{r^2\Lambda^2}{m}} + M\sqrt{\frac{\log\frac{1}{\delta}}{2m}} \text{ for all } h \in \mathcal{H}.$$

Proof. We define hypothesis sets $\mathcal{H}' \triangleq \{x \mapsto h(x) - y : h \in \mathcal{H}\}$ and $\mathcal{H}_\epsilon \triangleq \{x \mapsto |h(x) - y|_\epsilon : h \in \mathcal{H}\}$. Each hypothesis in $\mathcal{H}_\epsilon$ is 1-Lipschitz, and the result follows from Theorem 1.6 and bound on the empirical Rademacher complexity of $\mathcal{H}$. □

2.3 Lasso

Definition 2.6 (Least Absolute Shrinkage and Selection Operator (LASSO)). We consider the input space $\mathcal{X} = \mathbb{R}^N$, a hypothesis set of linear functions

$$\mathcal{H} \triangleq \left\{ x \mapsto \langle \mathbf{w}, x \rangle : \mathbf{w} \in \mathbb{R}^N \right\},$$

and the optimal hypothesis is defined as

$$\arg\min_{\mathbf{w},b} F(\mathbf{w}) = \arg\min_{\mathbf{w}} \lambda\|\mathbf{w}\|_1 + C\sum_{i=1}^{m} (y_i - \langle \mathbf{w}, x_i \rangle)^2.$$

Remark 10. LASSO problem defers from kernel ridge regression and support vector regression in that the complexity term is taken as L^1 norm rather than L^2 norm of weight vector $\mathbf{w}$. The empirical error for LASSO and kernel ridge regression is L^2 norm of the prediction error, whereas it is modified L^1 norm for support vector regression.

Remark 11. Taking complexity term as L^1 norm ensures a sparse solution for the weight vector $\mathbf{w}$. To achieve sparsity, one would like to minimize the L^0 (which is not a norm). However, this is a combinatorial task, since this is equivalent to the support of $\mathbf{w}$. The L^1 norm is a convex surrogate for L^0.

Remark 12. For this algorithm, we can show that for a sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ with $\|x_i\|_\infty \leqslant r_\infty < \infty$ for all $i \in [m]$, set of linear hypotheses $\mathcal{H} \triangleq \{x \mapsto \langle \mathbf{w}, x \rangle : \|\mathbf{w}\|_1 \leqslant \Lambda_1\}$, the empirical Rademacher complexity is upper bounded as

$$\hat{\mathfrak{R}}_m(\mathcal{H}) \leqslant \sqrt{\frac{2r_\infty^2 \Lambda_1^2 \log 2N}{m}}$$

In addition, if $|c(x)| \leqslant \Lambda_1 r_\infty$, then by triangular inequality and Hölder's inequality applied to conjugate pair $(1, \infty)$, we obtain

$$|h(x) - c(x)| \leqslant |h(x)| + |c(x)| \leqslant \|\mathbf{w}\|_1 \|x\|_\infty + \Lambda_1 r_\infty \leqslant 2\Lambda_1 r_\infty.$$

Theorem 2.7. *Consider a sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ with $\|x_i\|_\infty \leqslant r_\infty < \infty$ for all $i \in [m]$, set of linear hypotheses $\mathcal{H} \triangleq \{x \mapsto \langle \mathbf{w}, x \rangle : \|\mathbf{w}\|_1 \leqslant \Lambda_1\}$, and $|c(x)| \leqslant \Lambda_1 r_\infty$. Then for a fixed $\delta > 0$, with probability at least $1 - \delta$, we have*

$$R(h) \leqslant \hat{R}(h) + \frac{8r_\infty^2 \Lambda_1^2}{\sqrt{m}} \Big[\sqrt{\log 2N} + \frac{1}{2} \sqrt{\frac{\log \frac{1}{\delta}}{2}} \Big].$$

Remark 13. On the plus side, Lasso is (a) a quadratic optimization problem with convex constraints, (b) thus has efficient schemes to find the optimal weight vector for the given optimization problem, and has (c) strong theoretical generalization bounds. On the flip side, it has (a) no closed form solution, and (b) can not be used readily with kernels.