1 Techniques that Handle Overfitting

- Cross Validation:
  Hold out part of the training data and use it as a proxy for the generalization error
  Disadvantages: 1. Wastes data. 2. Time-consuming because a lot of the variants of
  cross validation involve doing multiple splits on data for training and validation and
  running the algorithm multiple times.

- Structural Risk Minimization:
  Earlier, we found an upper bound on the generalization error in the following form.
  Under usual assumptions, with probability at least $1 - \delta$, $\forall h \in \mathcal{H}$ and $|\mathcal{H}| < \infty$,
  $$err(h) \leq \hat{err}(h) + O\left(\sqrt{\frac{\ln|\mathcal{H}| + \ln(\frac{1}{\delta})}{m}}\right)$$
  This technique tries to minimize the entire right-hand side of the inequality.

- Regularization
  This general family of techniques is closely related to structural risk minimization.
  It minimizes expressions of the form $\hat{err} + \text{constant} \times \text{“complexity”}$

- Algorithms that tend to resist overfitting

2 Rademacher Complexity

We have already learned about using the growth function and VC-dimension as complexity
measures for infinite hypothesis spaces. Today, we are going to introduce a more modern and
elegant complexity measure called the Rademacher complexity. This technique subsumes
the previous techniques in the sense that the previous bounds we found using $|\mathcal{H}|$, the
growth function or the VC-dimension would fall out as special cases of the new measure.

2.1

We start by laying down the setups of Rademacher complexity.

Sample $S = \langle (x_1, y_1), ..., (x_m, y_m) \rangle$, $y_i \in \{-1, 1\}$. We are using $\{-1, 1\}$ here instead of
$\{0, 1\}$, in order to make the math come out nicer.

hypothesis $h : X \to \{-1, 1\}$
Here, we’re providing an alternative definition for training error.

$$\hat{err}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x_i) \neq y_i\}$$ (1)
\[ e\hat{r}(h) = \frac{1}{m} \sum_{i=1}^{m} \frac{1 - y_i h(x_i)}{2} = \frac{1}{2} - \frac{1}{2m} \sum_{i=1}^{m} y_i h(x_i) \] (2)

Equation (2) is reached because \( y_i h(x_i) \) equals 1 when \( y_i = h(x_i) \) and \( y_i h(x_i) \) equals \(-1\) when \( y_i \neq h(x_i) \).

\[ \frac{1}{m} \sum_{i=1}^{m} y_i h(x_i) = 1 - 2e\hat{r}(h) \] (3)

Training error is a reasonable measure of how well a single hypothesis fits the data set. From equation (3), we can see that in order to minimize the training error, we can simply maximize \( \frac{1}{m} \sum_{i=1}^{m} y_i h(x_i) \).

2.2

Now, let us introduce a random label for data \( i \), which we name \( \sigma_i \) and which is also known as a Rademacher random variable.

\[ \sigma_i = \begin{cases} -1, & \text{with probability 1/2} \\ +1, & \text{with probability 1/2} \end{cases} \] (4)

We can use this random label to form a complexity measure for \( \mathcal{H} \) that is independent of the real labels of \( \mathcal{S} \).

\[ E_\sigma[\max_{h \in \mathcal{H}} \frac{1}{m} \sum_{i} \sigma_i h(x_i)] \] (5)

Equation (5) intuitively measures the complexity of \( \mathcal{H} \). Notice that we can find the range of this measure using two extreme cases.

- \( \mathcal{H} = \{h_0\} \): because there is only one hypothesis, max is not used. We then arrive at the expectation of 0.
- \( \mathcal{S} \) is shattered by \( \mathcal{H} \): In this case, we can always find a hypothesis that matches all \( \sigma_i \). Thus, the expected value is 1.

We now know that this measure ranges from 0 to 1.
2.3

We now replace \( \mathcal{H} \) with \( \mathcal{F} \), a family of functions \( f : \mathcal{Z} \to \mathcal{R} \). This generalizes our hypotheses to real-valued functions.

Sample \( \mathcal{S} = \langle z_1, ..., z_m \rangle \), where \( z_i \in \mathcal{Z} \).

The definition for the empirical Rademacher complexity is

\[
\hat{R}_S(\mathcal{F}) = E_{\sigma} \left[ \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right]
\]

Notice we replaced max by sup (supremum) because max might not exist when taken over an infinite number of functions. Supremum takes the least upper bound. For example, \( \sup \{0.9, 0.99, 0.999, ... \} = 1 \).

In order to find a measure with respect to the distribution \( \mathcal{D} \) over \( \mathcal{Z} \), we take the expected value of the empirical Rademacher complexity and arrive at the definition for the expected Rademacher complexity — equation (7).

\[
\mathcal{R}_m(\mathcal{F}) = E_{\mathcal{S}} \left[ \hat{R}_S(\mathcal{F}) \right]
\]

\( \mathcal{S} = \langle z_1, ..., z_m \rangle \), where \( z_i \sim \mathcal{D} \)

3 Generalization Bounds Based on Rademacher Complexity

Theorem

Let \( \mathcal{F} \) be a family of functions \( f : \mathcal{Z} \to [0, 1] \). Assume \( \mathcal{S} = \langle z_1, ..., z_m \rangle \), i.i.d and \( z_i \sim \mathcal{D} \). Define \( \hat{E}_S[f] = \frac{1}{m} \sum_i f(z_i) \), \( E[f] = E_{z \sim \mathcal{D}}[f(z)] \). (\( \hat{E}_S[f] \) is similar to the idea of the training error and \( E[f] \) is similar to the idea of the generalization error)

With probability at least \( 1 - \delta \), \( \forall f \in \mathcal{F} \),

\[
E[f] \leq \hat{E}_S[f] + 2\mathcal{R}_m(\mathcal{F}) + O\sqrt{\frac{\ln(\frac{1}{\delta})}{m}}
\]

(8)

\[
E[f] \leq \hat{E}_S[f] + 2\hat{R}_S(\mathcal{F}) + O\sqrt{\frac{\ln(\frac{1}{\delta})}{m}}
\]

(9)

Proof

We want to bound the following random variable:

\[
\Phi(\mathcal{S}) = \sup_{f \in \mathcal{F}} (E[f] - \hat{E}_S[f])
\]

(10)
Step 1

Using the definitions, we get:

\[
\Phi(S) = \sup_{f \in \mathcal{F}} (E[f] - \hat{E}_S[f]) = \sup_{f \in \mathcal{F}} (E[f] - \frac{1}{m} \sum_i f(z_i)) \tag{11}
\]

Since \(f(z_i) \in [0, 1]\), changing any \(z_i\) value to \(z_i'\) can only change \(\frac{1}{m} \sum_i f(z_i)\) by at most \(\frac{1}{m}\), and therefore \(\Phi(S)\) by at most \(\frac{1}{m}\). This means that \(\Phi(S)\) satisfies the condition for McDiarmid’s inequality, in that \(|\Phi(z_1, ..., z_i, ..., z_m) - \Phi(z_1, ..., z_i', ..., z_m)| \leq c_i\), where \(c_i = \frac{1}{m}\).

McDiarmid’s inequality states that with probability at least \(1 - \delta\)
\[
Pr[f(x_1, ..., x_m) - E[f(X_1, ..., X_m)] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{2m} c_i^2}\right)
\]

Applying McDiarmid’s inequality, we get:
With probability at least \(1 - \delta\)
\[
\Phi(S) \leq E_S[\Phi(S)] + \sqrt{\frac{\ln(\frac{1}{\delta})}{2m}} \tag{12}
\]

Step 2

Let us define a ghost sample \(S' = \langle z_1', ..., z_m' \rangle\), \(z_i' \sim \mathcal{D}\). We aim to show that \(E[\Phi(S)] \leq E_{S,S'}[\sup_{f \in \mathcal{F}} (\hat{E}_{S'}[f] - \hat{E}_S[f])]\).

\[
E_{S'}[\hat{E}_{S'}[f]] = E[f] \tag{13}
\]
Equation (13) is true because the expected value of the random variable \(\hat{E}_{S'}[f]\) over all samples \(S'\) is \(E[f]\).

\[
E_S[\hat{E}_S[f]] = \hat{E}_S[f] \tag{14}
\]
Equation (14) is true because the random variable \(\hat{E}_S[f]\) is independent of \(S'\).

Therefore,
\[
E[\Phi(S)] = E_S[\sup_{f \in \mathcal{F}} (E[f] - \hat{E}_S[f])] = E_S[\sup_{f \in \mathcal{F}} (E_S[\hat{E}_S[f]] - \hat{E}_S[f])] \leq E_{S,S'}[\sup_{f \in \mathcal{F}} (\hat{E}_{S'}[f] - \hat{E}_S[f])]
\]

The last inequality is true because the expected value of the max of some function is at least the max of the expected value of the function.
Step 3

Continuing the ghost sampling technique, we now try to obtain two new samples $T$ and $T'$ by running through the following mechanism on $S$ and $S'$.

for $i = 1, \ldots, m$
  with probability $1/2$: swap $z_i, z'_i$
  else: leave alone

$T, T' = \text{resulting samples}$

\[
E_{T'}[f] - E_T[f] = \frac{1}{m} \sum_i \begin{cases} (f(z_i) - f(z'_i)), & \text{with probability } 1/2 \\ (f(z'_i) - f(z_i)), & \text{with probability } 1/2 \end{cases} \tag{15}
\]

\[
\Rightarrow \quad E_{T'}[f] - E_T[f] = \frac{1}{m} \sum_i \sigma_i (f(z'_i) - f(z_i)) \tag{16}
\]

We know that $T, T' \sim S, S'$ (equally distributed) because $S, S'$ are i.i.d samples from the distribution $D$.

Therefore, $\sup_{f \in \mathcal{F}} (E_{S'}[f] - E_S[f]) \sim \sup_{f \in \mathcal{F}} (\frac{1}{m} \sum_i \sigma_i (f(z'_i) - f(z_i)))$.

Then, if we take the expected values of the two expressions over $S, S'$ and $\sigma_i$, the values should equal to each other.

Equation (17) shows the conclusion for step 3.

\[
E_{S, S'}[\sup_{f \in \mathcal{F}} (E_{S'}[f] - E_S[f])] = E_{S, S', \sigma}[\sup_{f \in \mathcal{F}} (\frac{1}{m} \sum_i \sigma_i (f(z'_i) - f(z_i)))] \tag{17}
\]