12. Structural Risk Minimization
ECE 830 & CS 761, Spring 2016
General setup for statistical learning theory

We observe training examples \( \{x_i, y_i\}_{i=1}^n \)

\[ x_i = \text{features } \in \mathcal{X} \]
\[ y_i = \text{labels / responses } \in \mathcal{Y} \]

**Definition: predictor / classifier**

A predictor is a function \( f : \mathcal{X} \mapsto \mathcal{Y} \), where \( \mathcal{X} \) is the feature space and \( \mathcal{Y} \) is the label space. Let \( \mathcal{F} \) be a collection of predictors.

We also have a loss function \( \ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_+ \). E.g. \( y \) is true label, \( \hat{y} \) is predicted label, and we measure \( \ell(y, \hat{y}) \geq 0 \). In this lecture we focus on \( \ell(y, \hat{y}) = 1_{\{y \neq \hat{y}\}} \) (0/1 loss).

Main assumption: \( (x_i, y_i) \overset{iid}{\sim} P \) iid, where \( P \) is unknown.
Goal:

Select an \( f \in \mathcal{F} \) so that it minimizes

\[
\text{Risk} = R(f) = \mathbb{E}_{(x,y) \sim \mathcal{P}}[\ell(y, f(x))]
\]

So far we have focused on empirical risk minimization, where

\[
\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))
\]

is the empirical risk. Then the empirical risk minimizer (ERM) is

\[
\hat{f} = \arg \min_{f \in \mathcal{F}} \hat{R}(f).
\]
What can be said about the ERM’s performance?

**Case 1: Finite sets of classifiers.** We showed that with probability $\geq 1 - \delta$

$$R(f) \lesssim \hat{R}(f) + \sqrt{\frac{\log |\mathcal{F}| + \log(1/\delta)}{n}} \quad \forall f \in \mathcal{F}$$

which leads to the bound

$$\mathbb{E}[R(\hat{f})] - \min_{f \in \mathcal{F}} R(f) \lesssim \sqrt{\frac{\log |\mathcal{F}| + \log n}{n}}$$
ERM performance

What can be said about the ERM’s performance?

**Case 2: Classifiers with finite VC dimension.** Let $S(\mathcal{F}, n)$ be the Shatter coefficient for $\mathcal{F}$, representing the number of different effective classifiers are in $\mathcal{F}$ for $n$ training samples; the VC dimension is

$$V_{\mathcal{F}} = \arg \max_{k \geq 1} \left\{ S(\mathcal{F}, k) = 2^k \right\}$$

and can be thought of as the largest number of examples that can be arbitrarily labeled. We showed using Sauer’s lemma, $S(\mathcal{F}, n) \leq (n + 1)^{V(\mathcal{F})}$ that with probability $\geq 1 - \delta$

$$|R(f) - \hat{R}(f)| \lesssim \sqrt{\frac{V \log n + \log(1/\delta)}{n}}$$

which leads to the bound

$$\mathbb{E}[R(\hat{f})] - \min_{f \in \mathcal{F}} R(f) \lesssim \sqrt{\frac{V \log n}{n}}.$$
Example: Histogram classifiers

Let $\mathcal{X} = [0, 1]^d$, $\mathcal{Y} = \{0, 1\}$ with 0/1 loss. Let $\mathcal{F}_k$ denote the set of histogram classifiers with $k$ bins. Note

$$|\mathcal{F}_k| = 2^k.$$ 

If we fix $m = k$ classifier bins, then we have with probability at least $1 - \delta$

$$R(f) - \hat{R}(f) \lesssim \sqrt{\frac{m \log 2 + \log(2/\delta)}{n}}.$$
Example: Histograms continued

Let $R^*$ be the Bayes’ Risk: $R^* = \min_f R(f)$ where the minimization is over all classifiers, including those potentially not in $\mathcal{F}$ and based on the unknown distribution $P$. Then for the empirical risk minimizer over $\mathcal{F} = \mathcal{F}_m$ we have

$$
\mathbb{E}[R(\hat{f})] - R^* = \mathbb{E}[R(\hat{f})] - \min_{f \in \mathcal{F}} R(f) + \min_{f \in \mathcal{F}} R(f) - R^* \\
\leq \sqrt{\frac{m \log(n)}{n}} + \min_{f \in \mathcal{F}} R(f) - R^*
$$

“estimation error”

“approximation error”

How should we choose $m$?

Choosing $m$ in advance based on $n$ means that we cannot optimally balance between the two terms in the bound for all distributions $P$. We might consider $\mathcal{F} = \bigcup_{k \geq 1} \mathcal{F}_k$, but this set has $|\mathcal{F}| = V(\mathcal{F}) = \infty$. 
Countably Infinite Sets of Classifiers

Suppose that $\mathcal{F}$ is a countable, possibly infinite, collection of candidate functions.

Example: Histogram classifiers

$$\mathcal{F} = \bigcup_{k \geq 1} \mathcal{F}_k$$

Further suppose that we have some prior distribution $p$ over this set, so that

$$p(f) \geq 0 \forall f \in \mathcal{F} \quad \text{and} \quad \sum_{f \in \mathcal{F}} p(f) = 1.$$ 

This provides two advantages:

1. By choosing $p(f)$ larger for certain $f$, we can preferentially treat those candidates
2. We do not need $\mathcal{F}$ to be finite and we only require $\sum_{f \in \mathcal{F}} p(f) = 1$
Let 

\[ c(f) = - \log p(f); \]

then we have

\[ \sum_{f \in \mathcal{F}} e^{-c(f)} = 1. \]

The numbers \( c(f) \) can be interpreted as

- log of prior probabilities
- codelengths
- measures of complexity
Now recall Hoeffding’s inequality. For each $f$ and every $\epsilon > 0$

$$\mathbb{P} \left( \mathcal{R}(f) - \hat{\mathcal{R}}_n(f) \geq \epsilon \right) \leq e^{-2n\epsilon^2}$$

or for every $\delta > 0$

$$\mathbb{P} \left( \mathcal{R}(f) - \hat{\mathcal{R}}_n(f) \geq \sqrt{\frac{\log(1/\delta)}{2n}} \right) \leq \delta$$

Suppose $\delta > 0$ is specified. Using the values $c(f)$ for $f \in \mathcal{F}$, define

$$\delta(f) := \delta e^{-c(f)}$$

Then we have

$$\mathbb{P} \left( \mathcal{R}(f) - \hat{\mathcal{R}}_n(f) \geq \sqrt{\frac{\log(1/\delta(f))}{2n}} \right) \leq \delta(f)$$
Furthermore we can apply the union bound as follows

\[
P \left( \bigcup_{f \in \mathcal{F}} R(f) - \hat{R}_n(f) \geq \sqrt{\frac{\log(1/\delta(f))}{2n}} \right) \leq \sum_{f \in \mathcal{F}} P \left( R(f) - \hat{R}_n(f) \geq \sqrt{\frac{\log(1/\delta(f))}{2n}} \right) \leq \sum_{f \in \mathcal{F}} \delta(f) = \sum_{f \in \mathcal{F}} e^{-c(f)} \delta = \delta
\]
We have that $\forall f \in \mathcal{F}$ and $\forall \delta > 0$ with probability at least $1-\delta$:

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{\log(1/\delta(f))}{2n}}$$

$$= \hat{R}_n(f) + \sqrt{\frac{c(f) + \log(1/\delta)}{2n}}$$
Example: Finite sets

Suppose \( \mathcal{F} \) is finite and \( c(f) = \log |\mathcal{F}| \) \( \forall f \in \mathcal{F} \) (this is a uniform prior). Then

\[
\sum_{f \in \mathcal{F}} e^{-c(f)} = \sum_{f \in \mathcal{F}} e^{-\log |\mathcal{F}|} = \sum_{f \in \mathcal{F}} \frac{1}{|\mathcal{F}|} = 1
\]

and

\[
\delta(f) = \frac{\delta}{|\mathcal{F}|}
\]

which implies \( \forall f \in \mathcal{F}, |\mathcal{F}| < \infty \), and \( \forall \delta > 0 \) with probability at least \( 1 - \delta \)

\[
R(f) \leq \hat{R}_n(f) + \sqrt{\frac{\log |\mathcal{F}| + \log(1/\delta)}{2n}}
\]

Note that this is precisely the PAC bound we derived in the last lectures.
Example: Histogram Classifiers

Let \( X = [0, 1]^d, Y = \{0, 1\} \). Let \( \mathcal{F}_k, k=1, 2, \ldots \) denote the collection of histogram classification rules with \( k \) equal volume bins, and let \( \mathcal{F} = \bigcup_{k \geq 1} \mathcal{F}_k \). For \( f \in \mathcal{F}_k \), we choose \( c(f) = 2^k \). (We will how to derive this and that it satisfies \( \sum_{f \in \mathcal{F}} e^{-c(f)} \leq 1 \) in the next lecture.)

Then \( \forall f \in \bigcup_{k \geq 1} \mathcal{F}_k \) and \( \forall \delta > 0 \), with probability at least \( 1 - \delta \)

\[
R(f) \leq \hat{R}_n(f) + \sqrt{\frac{2k_f \log 2 + \log(1/\delta)}{2n}}
\]

where \( k_f \) is the number of bins in histogram corresponding to \( f \).
Example: Histograms continued

Contrast with the bound we had for the class of $m$ bin histograms alone:

$$\forall f \in \mathcal{F}_m \text{ and } \forall \delta > 0, \text{ with probability } \geq 1 - \delta$$

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{m \log 2 + \log(1/\delta)}{2n}}$$

Notice the bound for all histograms rules is almost as good as the bound for only the $m$-bin rules. That is, when $k_f = m$ the bounds are within a factor of $\sqrt{2}$. On the other hand, the new bound is a big improvement, since it also gives us a guide for selecting the number of bins.
Beyond ERM

The above bounds can be used to derive a useful alternative to empirical risk minimization – one which exploits the prior encapsulated by $p(f) = e^{-c(f)}$. In general, we want to choose a classifier $\hat{f} \in \mathcal{F}$ so that

$$E[R(\hat{f}_n)] - \inf_{f \in \mathcal{F}} R(f)$$

is as small as possible. We’d like to choose

$$\hat{f} = \arg \min_{f \in \mathcal{F}} R(f)$$

but we cannot measure $R(f)$ to minimize it. However, we can minimize its upper bound!
**Definition: Structural risk minimizer**

For a countably infinite or finite class $\mathcal{F}$, let $c(f)$ be a function such that

$$\sum_{f \in \mathcal{F}} e^{-c(f)} \leq 1.$$

Then for any $\delta \in (0, 1)$,

$$\hat{f}_n^\delta = \arg \min_{f \in \mathcal{F}} \left\{ \hat{R}_n(f) + C(f, n, \delta) \right\}$$

where

$$C(f, n, \delta) \equiv \sqrt{\frac{c(f) + \log(2/\delta)}{2n}}$$

is the **structural risk minimizer**.
According to the PAC bound, $\forall f \in \mathcal{F}$ and $\forall \delta > 0$, with probability $\geq 1 - \delta$,

$$R(f) \leq \hat{R}_n(f) + C(f, n, \delta)$$

and in particular,

$$R(\hat{f}_n^\delta) \leq \hat{R}_n(\hat{f}_n^\delta) + C(\hat{f}_n^\delta, n, \delta)$$

so, by the definition of $\hat{f}_n^\delta$, $\forall f \in \mathcal{F}$

$$R(\hat{f}_n^\delta) \leq \hat{R}_n(f) + C(f, n, \delta)$$

We will make use of the inequality above in a moment. First note that $\forall f \in \mathcal{F}$

$$E[R(\hat{f}_n^\delta)] - R(f) = E[R(\hat{f}_n^\delta) - \hat{R}_n(f)] + E[\hat{R}_n(f) - R(f)]$$

The second term is exactly 0, since $E[\hat{R}_n(f)] = R(f)$. 
Now consider the first term \( E[R(\hat{f}_n^\delta) - \hat{R}_n(f)] \). Let \( \Omega \) be the set of events on which

\[
R(\hat{f}_n^\delta) \leq \hat{R}_n(f) - C(f,n,\delta), \quad \forall \ f \in \mathcal{F}
\]

From our PAC bound, we know that \( P(\Omega) \geq 1 - \delta \). Thus,

\[
E[R(\hat{f}_n^\delta) - \hat{R}_n(f)]
= E[R(\hat{f}_n^\delta) - \hat{R}_n(f)|\Omega]P(\Omega) + E[R(\hat{f}_n^\delta) - \hat{R}_n(f)|\Omega^c](1 - P(\Omega))
\leq C(f,n,\delta) + \delta \quad \text{(since } 0 \leq R, \ \hat{R} \leq 1, \ P(\Omega) \leq 1 \text{ and } 1 - P(\Omega) \leq \delta \)
\]

\[
= \sqrt{\frac{c(f) + \log(2/\delta)}{2n}} + \delta
\]

\[
= \sqrt{\frac{c(f) + \frac{1}{2} \log n}{2n}} + \frac{1}{\sqrt{n}} \quad \text{(by setting } \delta = \frac{1}{\sqrt{n}})\]
We can summarize our analysis with the following theorem.

**Theorem: Complexity Regularized Model Selection**

Let $\mathcal{F}$ be a collection of functions, and assign a positive number $c(f)$ to each $f \in \mathcal{F}$ such that $\sum_{f \in \mathcal{F}} e^{-c(f)} \leq 1$. Define the structural risk minimizer

$$
\hat{f}_n = \arg \min_{f \in \mathcal{F}} \left\{ \hat{R}_n(f) + \sqrt{\frac{c(f) + \frac{1}{2} \log n}{2n}} \right\}
$$

Then,

$$
E[R(\hat{f}_n)] \leq \inf_{f \in \mathcal{F}} \left\{ R(f) + \sqrt{\frac{c(f) + \frac{1}{2} \log n}{2n}} + \frac{1}{\sqrt{n}} \right\}
$$
This shows that

\[ \hat{R}_n(f) + \sqrt{\frac{c(f) + \frac{1}{2} \log n}{2n}} \]

is a reasonable surrogate for

\[ R(f) + \sqrt{\frac{c(f) + \frac{1}{2} \log n}{2n}} \]
Example: Histogram Classifiers

Let $\mathcal{X} = [0, 1]^d$ be the input space and $\mathcal{Y} = \{0, 1\}$ be the output space. Let $\mathcal{F}_k$, $k = 1, 2, \ldots$ denote the collection of histogram classification rules with $k$ equal volume bins. Let $\hat{f}_n$ be the structural risk minimizer

$$\hat{f}_n = \arg \min_{f \in \mathcal{F}} \hat{R}(f) + \sqrt{\frac{c(f) + \frac{1}{2} \log n}{2n}}$$

Recall our choice $c(f) = 2k$ for $f$ a $k$-bin histogram classifier. Then equivalently

$$\hat{f}_n = \min_{k \geq 1} \left\{ \min_{f \in \mathcal{F}_k} \hat{R}_n(f) + \sqrt{\frac{2k + \frac{1}{2} \log n}{2n}} \right\}$$
Example: Histograms continued

That is, for each $k$, let

$$\hat{f}_n^{(k)} = \arg \min_{f \in F_k} \hat{R}_n(f)$$

Then select the best $k$ according to

$$\hat{k} = \arg \min_{k \geq 1} \left\{ \hat{R}_n(\hat{f}_n^{(k)}) + \sqrt{\frac{2k + \frac{1}{2} \log n}{2n}} \right\}$$

and set

$$\hat{f}_n = \hat{f}_n^{(\hat{k})}$$

Then,

$$E[R(\hat{f}_n)] \leq \inf_{k \geq 1} \left\{ \min_{f \in F_k} R(f) + \sqrt{\frac{2k + \frac{1}{2} \log n}{2n}} + \frac{1}{\sqrt{n}} \right\}$$
From the third homework, we know that if $d = 2$ and the Bayes decision boundary is a 1-d curve, then by setting $k = \sqrt{n}$ and selecting the best $f$ from $\mathcal{F}_{\sqrt{n}}$ we have

$$E[R(\hat{f}_n)] = O(n^{-1/4})$$

It is a simple exercise to show that the complexity regularized classifier will perform just as well automatically. That is, the proper $k$ is selected automatically, without user intervention.