ECE 0402 - Pattern Recognition

LECTURE 2

Today (1/19): We will talk about: why it is important to do inference from data using statistical perspective, and have a first look at the theory of generalizations for the binary supervised calssification problem.

Supervised Learning: Given training data $(x_1, y_1), ..., (x_n, y_n)$, we would like to learn a function $f: X \mapsto Y$ such that $f(x) = y$ for x other than $x_1, ..., x_n$

Without any additional assumptions, we conclude nothing about f except for its value on this finite set inputs $x_1, ..., x_n$ – NOT so correct!

Probabilistic perspective: Generalization

- Any f agreeing with the training data may be **possible**.
- But that doesn't mean that any f is equally **probable**

  Example: $P[heads] = p,\ P[tails] = 1 - p$ Biased coin

  - toss the coin n times (independently)
  - $\hat{p} = \frac{\#of\ heads}{n}$
  - does $\hat{p}$ tell us anything about p?
  - for large n we expect $\hat{p} \approx p$ $\hat{p}$ is a good estimate as $n \to \infty$
  - Law of large of numbers:

$$\hat{p} \to p \quad as \quad n \to \infty$$

  - we can learn something about p from observations– at least in a very limited sense
  - there is always the **possibility** that we are totally wrong ($\hat{p} \neq p$), but given enough data, the **probability** should be very small

coin tosses: we want to estimate p

learning: we want to estimate a function $f: X \mapsto Y$

Suppose we have hypothesis h - guess for f

and think of the (x_i, y_i) as series of independent coin tosses where (x_i, y_i) are drawn from a probability distribution

- heads: our hypothesis is correct, i.e., $h(x_i) = y_i$

- tails: our hypothesis is wrong, i.e., $h(x_i) \neq y_i$

Probability that it disagrees

Definition:

(True) Risk : $R(h) := \mathbb{P}[h(X) \neq Y]$

Indicator function

Empirical Risk : $\hat{R}_n(h) := 1/n \sum_{i=1}^{n} 1_{\{h(x_i) \neq y_i\}}$ Access to points we have seen

The law of large numbers guarantees that as long as we have enough data, we will have that $R(h) \approx \hat{R}_n(h)$. This means that we can use $\hat{R}_n(h)$ to verify whether h was a good hypothesis

Empirical risk approximates true risk with enough data

- where did h come from?
- what if $R(h)$ is large?

Now consider, ensemble of many hypothesis $\mathcal{H} = h_1, ..., h_m$. If we fix h_j before drawing our data, then the LLN tells us that $\hat{R}_n(h_j) \to R(h_j)$. However, it is also true that for a fixed n, if m is large it can still be very likely that there is some hypothesis h_k for which $\hat{R}_n(h_k)$ is still very far from $R(h_k)$.

Example:

($P[H] = \frac{1}{2}$) (Independent trials) $\left(\frac{1}{2}\right)^{10}$

- Toss a fair coin 10 times, the probability of 10 heads: 0.001
- Toss 1000 fair coins 10 times, the probability that **some** coin will get 10 heads: 0.624

m is large

This phenomenon forms the fundamental challenge of multiple hypothesis testing. So the question is: how to adapt our approach to handle many hypothesis?

Assumption: There is some underlying function $f : X \mapsto Y$ that captures some input-output relationship that we would like to estimate. We do not know f, but we get to observe examples input-output pairs which are **generated independently at random**.

Two versions

- we draw x_i according to some unknown distribution and get to observe $(x_i, f(x_i))$
- x_i from some unknown distribution and we have $(x_i, f(x_i)+n_i)$ where n_i models "noise" with an unknown distribution – interpretation: the labels are not always going to be perfect
- we draw pairs (x_i, y_i) according to some unknown joint distribution

Example: Binary Classification Problem
As a start, let's focus on one example: Binary classification where you have two classes.

Ingredients:

- output: $Y = \{0, 1\}$ or $Y = \{-1, +1\}$ are the class labels.
- seen: the training data $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ where each $x_i \in \mathbb{R}^d$ are "feature vectors".
- The learning model consist of an algorithm and $\mathcal{H} = \{h_1, ..., h_m\}$ an ensemble of possible hypothesis–potential candidates rules for the **unknown** mapping of $X \mapsto Y$.
- algorithm select the "best" possible hypothesis from $\mathcal{H}$ set.

The data are assumed to be random – are independent samples generated from some joint distribution on $\mathbb{R}^d \times \{0, 1\}$, but we don't know anything about this distribution a priori.

Truly unknown distribution

Tools:

- Risk $R(h_j) := \mathbb{P}[h_j(X) \neq Y]$, in other words probability of error.
- Empirical Risk $\hat{R}_n(h_j) := 1/n \sum_{i=1}^{n} 1_{h_j(x_i) \neq y_i}(i)$

* Notation: Indicator function

$$1_{\{A\}}(t) = \begin{cases} 1 & \text{if } t \in A \\ 0 & \text{if } t \notin A \end{cases}$$

How to:

- Repeat: The empirical risk $\hat{R}_n(h_j)$ gives us an estimate of the true risk $\hat{R}(h_j)$, and from the LLN we know that $\hat{R}_n(h_j) \to R(h_j)$ as $n \to \infty$. Hence, this simple tool suggest the most natural way of learning in this framework.
- We have a set of hypothesis $\mathcal{H}$ and we want to choose one from that set to achieve a small risk, i.e., $h^* = \arg\min_{h_j \in \mathcal{H}} \hat{R}_n(h_j)$. Choose the hypothesis that minimizes the risk

  Not a bad strategy, known also as $\boxed{\textbf{Empirical Risk Minimization (ERM)}}$

  But what if $\mathcal{H}$ is a big set, twenty trillion hypothesis, or even infinite? You may not wanna actually compute the empirical risk. We will have to do something else to search over this...

Don't want to actually compute empirical risk for every hypothesis

Side effects/danger:

- Is it a good idea to go after ERM?
  - If we have enough data, large n; LLN tells us empirical risk is a good estimate of true risk, $\hat{R}_n(h_j) \approx R(h_j)$.
  - However, we also have this other problem that says: if the number of hypothesis m is very large , then maybe one of these h_k's in this set will give $\hat{R}_n(h_k) \ll R(h_k)$ or $\hat{R}_n(h_k) \gg R(h_k)$.

$h^* \leftarrow h_1, \ldots, h_m$

- What can we say about $R(h^*)$?
  - we know $\hat{R}_n(h^*)$ is small, this could be because true risk $R(h^*)$ was small. OR...
  - it was one those examples where the empirical risk was much smaller than true risk $\hat{R}_n(h_k) \ll R(h_k)$ for some h_k.

  How do we know which one of these would have generated it? Which explanation is these two we think is more likely, vote? Depends on n and m

  - If we are deciding between m hypothesis, how much data we need to ensure that

$$| \hat{R}_n(h^*) - R(h^*) | \leq \epsilon \quad (*)$$

    for some $\epsilon \in (0, 1)$.
  - If we want a result like this, we can't use asymptotic results like LLN and CLT do not give us answers to these questions.
  - We want a non-asymptotic result, we wanna be able to quantify if m is finite and n is finite, what is the probability that (*) holds.

$$\mathbb{P}\Big[\ | \hat{R}_n(h^*) - R(h^*) | \leq \epsilon \Big]$$

Our goal boils down to make

$$\mathbb{P}\Big[\ | \hat{R}_n(h^*) - R(h^*) | \leq \epsilon \Big] \approx 1 \quad (**)$$

by setting n appropriately. What is random in this statement?

- $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ Data
- $\hat{R}_n(h_1), ..., \hat{R}_n(h_m)$ Risk (depends on data)
- h^* Hypothesis (depends on data)

$\hat{R}(n)$ is empirical risk

$R(n)$ is true risk

Let's just analyze a single hypothesis to make the problem easier.

$$\mathbb{P}[\ \mid \hat{R}_n(h_j) - R(h_j) \mid \leq \epsilon\ ] \approx 1$$

Now that h_j is just fixed, true risk for it is a number, and $\hat{R}_n(h_j)$ is the only random entity in here.

We can write empirical risk of hypothesis h_j as a sum of Bernoulli random variables (RVs):

Sum of Bernoulli RVs is binomial RV

$$\hat{R}_n(h_j) := 1/n \sum_{i=1}^{n} 1_{h_j(x_i) \neq y_i} = 1/n \sum_{i=1}^{n} S_i$$

S_i is Bernoulli RVs, thus, $n\hat{R}_n(h_j)$ is a Binomial RV. Since $\mathbb{P}[S_i = 1] = \mathbb{P}[h_j(x_i) \neq y_i] = R(h_j)$

(error) $\rightarrow$ true risk

(Expectation of sum is sum of expectations)

$$E\left[n\hat{R}_n(h_j)\right] = E\left[\sum_{i=1}^{n} S_i\right] = \sum_{i=1}^{n} E[S_i]$$
$$= n\ \mathbb{P}[h_j(x_i) \neq y_i]$$
$$= n\ R(h_j)$$

This give us an equivalent way of thinking about our problem,

$$\mathbb{P}\left[\mid n\hat{R}_n(h_j) - nR_n(h_j) \mid\ \ \leq\ n\epsilon\right]$$

This is the probability that a Binomial RV will deviate from its mean by more than $n\epsilon$.

$$\mathbb{P}\left[\ \mid n\hat{R}_n(h_j) - nR_n(h_j) \mid\ \ \leq\ n\epsilon\right] = F(nR(h_j) + n\epsilon) - F(nR(h_j) - n\epsilon)$$

If we remember the CDF of a binomial Rv:

$$F(a) = \sum_{i=0}^{\lfloor a \rfloor} \binom{n}{i} R(h_j)^i (1 - R(h_j))^{(n-i)}$$

Rather than calculating this probability exactly, it is good enough to get a good bound on it, i.e. looking for an inequality of the form:

$$\mathbb{P}\left[\mid \hat{R}_n(h_j) - R(h_j) \mid\ \ \leq\ \epsilon\right] \geq\ 1 - ?$$

or equivalently,

$$\mathbb{P}\left[\mid \hat{R}_n(h_j) - R(h_j) \mid\ \ \geq\ \epsilon\right] \leq\ ?$$

- **Markov's Concentration Inequality:** for $X \geq 0$ any nonnegative RV, and any $t \geq 0$:

$$\mathbb{P}\left[X \geq t\right] \leq \frac{\mathbb{E}[X]}{t}$$

From Markov's Inequality, for any strictly monotonically increasing (non-negative-valued) function ϕ:

$$\mathbb{P}[X \geq t] = \mathbb{P}\left[\phi(X) \geq \phi(t)\right] \leq \frac{\mathbb{E}\left[\phi(X)\right]}{\phi(t)}$$

The first result is Chebyshev's Inequality.

- **Chebyshev's Inequality:**

$$\mathbb{P}\left[\ \mid X - \mathbb{E}[X] \mid \ \geq \ \epsilon\right] \leq \frac{var[X]}{\epsilon^2}$$

(You can prove this using Markov's. Markov's inequality's proof is also straight forward, They are super useful, worth to review if you forgot...)

- **Hoeffding's Inequality:** This is the most useful one for our case. It assumes a bit more about our RV beyond having a finite variance, but gets us a much tighter bound.

Let $X_i, ..., X_n$ be independent bounded RVs (more assumption, we are not just talking about non-negative random variables, bounded in interval), $\mathbb{P}[X_i \in [a,b]] = 1$ for all i. Let $S_n = \sum_{i=1}^{n} X_i$. Then for any $\epsilon > 0$, we have;

$$\mathbb{P}\left[\mid S_n - \mathbb{E}[S_n] \mid \ \geq \ \epsilon\right] \leq 2\ e^{-\frac{2\epsilon^2}{n(b-a)^2}}$$

If you are trying to bound $\mathbb{P}[\mid S_n - \mathbb{E}[S_n] \mid \geq \epsilon]$, there is two ways you could violate it. You could have S_n is too big or S_n is too small. To begin consider only the upper tail inequality:

$$\begin{aligned}
\mathbb{P}[S_n - \mathbb{E}[S_n] \geq \ \epsilon] &= \mathbb{P}[\lambda\ S_n - \mathbb{E}[S_n] \geq \ \lambda\epsilon] \quad (\lambda > 0) \\
&= \mathbb{P}[e^{\lambda\ (S_n - \mathbb{E}[S_n])} \geq \ e^{\lambda\epsilon}] \ \text{, apply Markov Ineq. to this} \\
&\leq \frac{e^{\lambda\ (S_n - \mathbb{E}[S_n])}}{e^{\lambda\epsilon}} \\
&= e^{-\lambda\epsilon}\ \ \mathbb{E}\left[\ e^{\lambda(X_1 - \mathbb{E}[X_1] + ... + X_n - \mathbb{E}[X_n])}\ \right] \\
&= e^{-\lambda\epsilon}\ \ \prod_{i=1}^{n} \mathbb{E}\ \left[\ e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \quad \text{independence}
\end{aligned}$$

Using Hoeffding's Lemma, it is not obvious but also not too hard to show that (to prove use convexity and then get a bound using Taylor series expansion),

$$\mathbb{E}\ \left[\ e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \leq e^{\lambda^2 (b-a)^2/8}$$

Plugging this in, we obtain that for any positive λ,

$$\mathbb{P}[S_n - \mathbb{E}[S_n] \geq \epsilon] \ \leq \ e^{-\lambda\epsilon} e^{\lambda^2 (b-a)^2/8}$$

By setting $\lambda = \frac{4\epsilon}{n(b-a)^2}$,

$$\begin{aligned}\mathbb{P}[S_n - \mathbb{E}[S_n] \geq \epsilon] \leq \ & e^{-\frac{4\epsilon^2}{n(b-a)^2}} e^{\frac{2\epsilon^2}{n(b-a)^2}} \\ &= e^{-\frac{2\epsilon^2}{n(b-a)^2}}\end{aligned}$$

Okay, this is for S_n too big, bounded! You can use the same argument and do the other version (lower tail probability):

$$\mathbb{P}[\mathbb{E}[S_n] - S_n \geq \epsilon] = e^{-\frac{2\epsilon^2}{n(b-a)^2}}$$

Finally, combined:

$$\mathbb{P}[\mid S_n - \mathbb{E}[S_n] \mid \geq \epsilon] \leq 2e^{-\frac{2\epsilon^2}{n(b-a)^2}}$$

Special case: X_i are Bernoulli RVs, then S_n is a Binomial RV, and Hoeffding's Inequality becomes:

$$\mathbb{P}[\mid S_n - \mathbb{E}[S_n] \mid \geq \epsilon] \ \leq \ 2e^{-\frac{2\epsilon^2}{n}}$$

Going back to our original problem, we are interested in:

$$\mathbb{P}[\mid \hat{R}_n(h_j) - R(h_j) \mid \ \geq \ \epsilon]$$

This is not exactly Binomial, we need to multiply with n, and use Hoeffding's Lemma:

$$\begin{aligned}&\mathbb{P}[\mid \hat{R}_n(h_j) - R(h_j) \mid \ \geq \ \epsilon] \\ &\qquad = \mathbb{P}[\mid n \ \hat{R}_n(h_j) - n \ R(h_j) \mid \ \geq \ n\epsilon] \\ &\qquad \leq 2 \ e^{-2\epsilon^2 n}\end{aligned}$$

As n gets really big, bounds gets tighter exponentially fast! Strong statement. Thus after much effort, we have that for "a particular" hypothesis h_j,

$$\mathbb{P}[\mid \hat{R}_n(h_j) - R(h_j) \mid \ \geq \ \epsilon] \ \leq \ 2 \ e^{-2\epsilon^2 n}$$

However, we are ultimately interested in h^*, not just a single hypothesis h_j.
One way to argue that $\mid \hat{R}_n(h^*) - R(h^*) \mid \ \leq \ \epsilon$ is to ensure that $\mid \hat{R}_n(h_j) - R(h_j) \mid \ \leq \ \epsilon$ **simultaneously** for all possible j. We can express this mathematically as

$$\begin{aligned}\mathbb{P}[\ \mid \hat{R}_n(h^*) - R(h^*) \mid \geq \epsilon] \leq \mathbb{P}[\ & \mid \hat{R}_n(h_1) - R(h_1) \mid \geq \epsilon \\ \text{OR} \ & \mid \hat{R}_n(h_2) - R(h_2) \mid \geq \epsilon \\ & \vdots \\ \text{OR} \ & \mid \hat{R}_n(h_m) - R(h_m) \mid \geq \epsilon \ ]\end{aligned}$$

Union Bound

Generalization of m

By that, we can show:

$$\mathbb{P}[\ \mid \hat{R}_n(h^*) - R(h^*) \mid \geq \epsilon] \leq\ 2m\ e^{-2\epsilon^2 n}$$