CS229 lecture 4 notes

James Chuang

May 1, 2017

Contents

My notes on Andrew Ng's CS229 lecture 4 notes.

Learning Theory

1. Bias/variance tradeoff

- also see ESL Ch 2.9 and Ch 7
- is a more complex/flexible/high-capacity model better than a simple/inflexible/low-capacity model?
- some informal definitions:
    - ***generalization error***: the expected error on samples not necessarily in the training set
    - ***bias***: the expected generalization error even if a model were fit to a very (infinitely) large training set
        - high bias corresponds with ***underfitting***: i.e. failing to capture structure exhibited by the data
    - ***variance***: how much the generalization error is expected to change if the training set changes
        - high variance corresponds with ***overfitting***: i.e. fitting to the noise in the training set
    - there is a ***bias-variance tradeoff***:
        - a simple/inflexible/low-capacity model with few parameters may have large bias (but smaller variance)
        - a complex/flexible/high-capacity model with many parameters may have large variance (but smaller bias)

2. Preliminaries

- things we want to do:
    1. make the bias/variance tradeoff formal
        - this will lead to model selection methods, e.g. for choosing what order polynomial to fit to a training set
    2. relate error on the training set to generalization error
        - we care about generalization error, but we train models on training sets
    3. find conditions under which we can prove that learning algorithms will work well?
- two simple but useful lemmas:
    - ***the union bound***
        - Let $A_1, A_2, \ldots, A_k$ be k different (not necessarily independent) events. Then,

$$P\left(A_1 \cup \cdots \cup A_k\right) \leq P(A_1) + \cdots + P(A_k).$$

        - in words, the probability of any one of k events happening is at most the sums of the probabilities of the k different events
    - ***Hoeffding inequality*** aka the ***Chernoff bound*** in learning theory
        - Let $Z_1, \ldots, Z_m$ be m i.i.d. random variables drawn from a Bernoulli(ϕ) distribution, i.e.

$$P(Z_i = 1) = \phi \quad \text{and} \quad P(Z_i = 0) = 1 - \phi$$

- Let $\hat{\phi} = \frac{1}{m} \sum_{i=1}^{m} Z_i$ be the mean of these random variables
- Let any $\gamma > 0$ be fixed. Then,

$$P\left(\left|\phi - \hat{\phi}\right| > \gamma\right) \leq 2 \exp\left(-2\gamma^2 m\right)$$

- in words, if we take $\hat{\phi}$ – the average of m Bernoulli(ϕ) random variables – to be our estimate of ϕ, then the probability of our being far from the true value is small, so long as m is large
  - note that this only applies to the case of mBernoulli random variables described here: the more general Hoeffding inequality is described in the supplemental notes
- in other words, if you have a biased coin whose chance of landing on heads is ϕ, then if you toss it m times and calculate the fraction of time that it came up heads, that will be a good estimate of ϕ with high probability (if m is large)

- first, restrict attention to binary classification with labels $y \in \{0, 1\}$
  - note that everything here generalizes to other problems, including regression and multi-class classification
  - assume a training set $S = \left\{\left(x^{(i)}, y^{(i)}\right); i = 1, \ldots, m\right\}$ of size m, where the training examples $\left(x^{(i)}, y^{(i)}\right)$ are drawn i.i.d. from some probability distribution $\mathcal{D}$
  - for a hypothesis h, define the ***training error*** (aka the ***empirical risk*** or ***empirical error*** in learning theory):

$$\hat{\mathcal{E}}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\left\{h\left(x^{(i)}\right) \neq y^{(i)}\right\}$$

  - i.e., the fraction of training examples that h misclassifies
    - when we want to make clear the dependence of $\hat{\mathcal{E}}(h)$ on the training set S, we can write it $\hat{\mathcal{E}}_S(h)$
  - define the generalization error to be:

$$\mathcal{E}(h) = P_{(x,y)\sim\mathcal{D}}\left(h(x) \neq y\right)$$

  - i.e., the probability that, if we draw a new example (x, y) from the distribution $\mathcal{D}$, it will be misclassified by h
    - note the assumption that the training data are drawn from the *same* distribution $\mathcal{D}$ with which the hypothesis is evaluated
      - this is sometimes referred to as one of the ***PAC*** (probably approximately correct) assumptions
  - consider the setting of linear classification
    - let $h_\theta(x) = 1\left\{\theta^T x \geq 0\right\}$
      - what's a reasonable way of fitting the parameters θ?
        - one approach: minimize the training error by picking:

$$\hat{\theta} = \arg\min_{\theta} \hat{\mathcal{E}}\left(h_\theta\right)$$

        - this is called ***empirical risk minimization*** (ERM)
          - the resulting hypothesis output by the learning algorithm is $\hat{h} = h_{\hat{\theta}}$
            - this is the most "basic" learning algorithm
  - in our study of learning theory, it will be useful to abstract away from the specific parameterization of hypothesis
    - define the ***hypothesis class*** $\mathcal{H}$ used by a learning algorithm to be the set of all classifiers considered by it
      - e.g., for linear classification, $\mathcal{H} = \left\{h_\theta : h_\theta(x) = 1\left\{\theta^T x \geq 0\right\}, \theta \in \mathbb{R}^{n+1}\right\}$ is the set of all classifiers over $\mathcal{X}$ (the domain of the inputs) where the decision boundary is linear
        - most broadly, if we were studying neural networks (for example), then $\mathcal{H}$ would be the set of all classifiers representable by some neural network architecture
    - empirical risk minimization is then a minimization over the class of functions $\mathcal{H}$, in which the learning algorithm picks the hypothesis:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\mathcal{E}}(h)$$

3. The case of finite $\mathcal{H}$

- Start by considering a learning problem with a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_k\}$ consisting of k hypotheses
  - $\mathcal{H}$ is a set of k functions mapping from $\mathcal{X}$ to $\{0, 1\}$
    - empirical risk minimization selects $\hat{h}$ to be whichever of these k functions has the smallest training error
      - we will derive some guarantees on the generalization error of $\hat{h}$:
        - first, we will show that $\hat{\mathcal{E}}(h)$ is a reliable estimate of $\mathcal{E}(h)$ for all h
          - second, we will show that this implies an upper-bound on the generalization error of $\hat{h}$
    - take any one, fixed $h_i \in \mathcal{H}$
      - consider a Bernoulli random variable Z whose distribution is defined as follows:
        - sample $(x, y) \sim D$
        - then, set $Z = 1\{h_i(x) \neq y\}$
          - i.e., draw one example, and let Z indicate whether h_i misclassifies it
        - similarly, define $Z_j = 1\left\{h_i\left(x^{(j)}\right) \neq y^{(j)}\right\}$
        - since the training set was drawn iid from $\mathcal{D}$, Z and the Z_j's have the same distribution
          - the misclassification probability on a randomly drawn example, i.e. $\mathcal{E}(h)$, is exactly the expected value of Z (and Z_j). Moreover, the training error can be written:

$$\hat{\mathcal{E}}(h_i) = \frac{1}{m}\sum_{j=1}^{m} Z_j$$

          - thus, $\hat{\mathcal{E}}(h_i)$ is exactly the mean of the m random variables Z_j that are drawn iid from a Bernoulli distribution with mean $\mathcal{E}(h_i)$
            - by the Hoeffding inequality:

$$P\left(\left|\mathcal{E}(h_i) - \hat{\mathcal{E}}(h_i)\right| > \gamma\right) \leq 2\exp(-2\gamma^2 m)$$

            - this shows that, for this particular h_i, training error will be close to generalization error with high probability, assuming m is large
              - to prove that this is simultaneously true for ***all*** $h \in \mathcal{H}$:
                - let A_i denote the event that $\left|\mathcal{E}(h_i) - \hat{\mathcal{E}}(h_i)\right|$
                - then, the above inequality (for a particular A_i) can be written $P(A_i) \leq 2\exp(-2\gamma^2 m)$
                - using the union bound:

$$\begin{aligned}
P\left(\exists\, h \in \mathcal{H}.\left|\mathcal{E}(h_i) - \hat{\mathcal{E}}(h_i)\right| > \gamma\right) &= P(A_1 \cup \cdots \cup A_k) \\
&\leq \sum_{i=1}^{k} P(A_i) \\
&\leq \sum_{i=1}^{k} 2\exp\left(-2\gamma^2 m\right) \\
&\leq 2k\exp\left(-2\gamma^2 m\right) \qquad \text{subtract both sides from } 1 \\
P\left(\neg\exists\, h \in \mathcal{H}.\left|\mathcal{E}(h_i) - \hat{\mathcal{E}}(h_i)\right| > \gamma\right) &\leq 1 - 2k\exp\left(-2\gamma^2 m\right) \\
P\left(\forall h \in \mathcal{H}.\left|\mathcal{E}(h_i) - \hat{\mathcal{E}}(h_i)\right| \leq \gamma\right) &\geq 1 - 2k\exp\left(-2\gamma^2 m\right)
\end{aligned}$$

                - i.e., with probability at least $1 - 2k\exp\left(-2\gamma^2 m\right)$, $\mathcal{E}(h)$ will be within γ of $\hat{\mathcal{E}}(h)$ for all $h \in \mathcal{H}$.
                  - this is a ***uniform convergence*** result because this bound holds simultaneously for *all* $h \in \mathcal{H}$.
      - what we did above was, given particular values of m and γ, put a bound on the probability that for some $h \in \mathcal{H}$, $\left|\mathcal{E}(h) - \hat{\mathcal{E}}(h)\right| > \gamma$
        - the three quantities of interest: m, γ, and the probability of error
          - each can be bounded in terms of the other two

    - e.g., we can ask, “Given γ and some $\delta > 0$, how large must m be before we can guarantee that with probability at least $1 - \delta$, training error will be within γ of generalization error?”

$$\begin{aligned}
1 - \delta &\geq 1 - 2k \exp\left(-2\gamma^2 m\right) \\
2k \exp\left(-2\gamma^2 m\right) &\geq \delta \\
\exp\left(-2\gamma^2 m\right) &\geq \frac{\delta}{2k} \\
-2\gamma^2 m &\geq \log \frac{\delta}{2k} \\
m &\leq \frac{1}{2\gamma^2} \log \frac{\delta}{2k} \\
m &\geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}
\end{aligned}$$

    - i.e., with probability at least $1 - \delta$, we have that $\left|\mathcal{E}(h) - \hat{\mathcal{E}}(h)\right| \leq \gamma \, \forall \, h \in \mathcal{H}$
        - equivalently, the probability $\left|\mathcal{E}(h) - \hat{\mathcal{E}}(h) > \gamma\right|$ for some $h \in \mathcal{H}$ is at most δ
        - this bound tells us how many training examples we need in order to make a guarantee
            - ***sample complexity***: the training set size m that an algorithm requires to achieve a certain level of performance
        - key property: the number of training examples needed to make this guarantee is only *logarithmic* in k, the number of hypotheses in $\mathcal{H}$
    - similarly, can hold m and δ fixed and solve for γ:

$$\begin{aligned}
-2\gamma^2 m &\geq \log \frac{\delta}{2k} \\
\gamma^2 &\leq -\frac{1}{2m} \log \frac{d}{2k} \\
\gamma^2 &\leq \frac{1}{2m} \log \frac{2k}{d} \\
\gamma &\leq \sqrt{\frac{1}{2m} \log \frac{2k}{d}} \\
\left|\hat{\mathcal{E}}(h) - \mathcal{E}(h)\right| &\leq \sqrt{\frac{1}{2m} \log \frac{2k}{d}}
\end{aligned}$$

    - assume that uniform convergence holds, i.e. $\left|\mathcal{E}(h) - \hat{\mathcal{E}}(h)\right| \leq \gamma \, \forall \, h \in \mathcal{H}$
        - what can we prove about the generalization of our learning algorithm that picked $\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\mathcal{E}}(h)$?
        - define $h^* = \arg\min_{h \in \mathcal{H}} \mathcal{E}(h)$ to be the best possible hypothesis in $\mathcal{H}$
            - h^* is the best hypothesis given $\mathcal{H}$, so it makes sense to compare performance relative to h^*:

$$\begin{aligned}
\left|\mathcal{E}\left(\hat{h}\right) - \hat{\mathcal{E}}\left(\hat{h}\right)\right| &\leq \gamma && && \\
\mathcal{E}\left(\hat{h}\right) &\leq \hat{\mathcal{E}}\left(\hat{h}\right) + \gamma && && \\
\mathcal{E}\left(\hat{h}\right) &\leq \hat{\mathcal{E}}\left(h^*\right) + \gamma && \hat{\mathcal{E}}\left(\hat{h}\right) \leq \hat{\mathcal{E}}\left(h^*\right) && \left|\mathcal{E}\left(h^*\right) - \hat{\mathcal{E}}\left(h^*\right)\right| \\
\mathcal{E}\left(\hat{h}\right) &\leq \mathcal{E}\left(h^*\right) + 2\gamma && && \hat{\mathcal{E}}\left(h^*\right) \leq \mathcal{E}\left(h^*\right) + \gamma
\end{aligned}$$

            - therefore, if uniform convergence occurs, then the generalization error of $\hat{h}$ is at most 2γ worse than the best possible hypothesis in $\mathcal{H}$!
- theorem:
    - Let $|\mathcal{H}| = k$
    - let m, δ be fixed
    - then, with probability at least $1 - \delta$:

$$\mathcal{E}\left(\hat{h}\right) \leq \left(\min_{h\in\mathcal{H}} \mathcal{E}\left(h\right)\right) + 2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

- this is proved by:
  1. letting γ equal the $\sqrt{\cdot}$ term
  2. the previous argument that uniform convergence occurs with probability at least $1-\delta$
  3. noting that uniform convergence implies that $\mathcal{E}(h)$ is at most 2γ higher than $\mathcal{E}(h^*) = \min_{h\in\mathcal{H}} \mathcal{E}(h)$
- this quantifies the bias/variance tradeoff in model selection
  - specifically, suppose we have some hypothesis class $\mathcal{H}$, and a much larger hypothesis class $\mathcal{H}' \supseteq \mathcal{H}$
  - if we choose $\mathcal{H}'$:
    - the first term $\min_{h\in\mathcal{H}}(h)$ can only decrease, so the bias can only decrease
    - k (the number of possible hypotheses) increase, so the second term $2\sqrt{\cdot}$ also increases, corresponding to an increase in variance
- by holding γ and δ fixed and solving for m as before, we also obtain the following sample complexity bound:
  - Let $|\mathcal{H}| = k$
  - let δ, γ be fixed
  - then, for $\mathcal{E}\left(\hat{h}\right) \leq \min_{h\in\mathcal{H}} \mathcal{E}(h) + 2\gamma$ to hold with probability at least $1-\delta$, it suffices that:

$$\begin{aligned} m &\geq \frac{1}{2\gamma^2}\log\frac{2k}{\delta} \\ &= O\left(\frac{1}{\gamma^2}\log\frac{k}{\delta}\right) \end{aligned}$$

4. The case of infinite $\mathcal{H}$

- many hypothesis classes contain an infinite number of functions
  - includes any parameterized by real numbers, e.g. linear classification
- first, an “incorrect” argument:
  - suppose we have $\mathcal{H}$ parameterized by d real numbers
    - a computer can only use a finite number of bits to represent an real number
      - IEEE double-precision floating point (i.e. a `double` in C) uses 64 bits to represent a floating point number
      - thus, the hypothesis class consists of at most $k = 2^{64d}$ different hypotheses
        - we therefore find that, to guarantee $\mathcal{E}\left(\hat{h}\right) \leq \mathcal{E}\left(h^*\right) + 2\gamma$ to hold with probability at least $1-\delta$, it suffices that:

$$\begin{aligned} &m \geq O\left(\frac{1}{\gamma^2}\log\frac{2^{64d}}{\delta}\right) \\ &m \geq O\left(\frac{d}{\gamma^2}\log\frac{1}{d}\right) \\ &m \geq O_{\gamma,\delta}\left(d\right) && O_{\gamma,\delta} \text{ indicates that } O \text{ is hiding constants dependent on } \gamma,\delta \end{aligned}$$

        - thus, the number of training examples needed is at most **linear** in the parameters of the model
        - this proof is not entirely satisfying since it relies on the precision of 64-bit floating point, but the conclusion is roughly correct: If trying to minimize training error, then in order to learn “well” using a hypothesis class that has d parameters, in general we need on the order of a linear number of training examples in d
          - note that this is proven for algorithms that use empirical risk minimization. Good theoretical guarantees on non-ERM learning algorithms are a subject of active research
        - this proof is also unsatisfying because it relies on the parameterization of $\mathcal{H}$
          - intuitively, the parameterization doesn't seem like it should matter
- in order to derive a more complete argument, we need a few definitions
  - Given a set $S = \left\{x^{(i)}, \ldots, x^{(d)}\right\}$ (unrelated to the definition of a training set) of points $x^{(i)} \in \mathcal{X}$:
    - we say that $\mathcal{H}$ ***shatters*** $\mathcal{S}$ if $\mathcal{H}$ can realize any labeling on S.

      - i.e., if for any set of labels $\left\{y^{(i)}, \dots, y^{(d)}\right\}$, there exists some $h \in \mathcal{H}$ so that $h\left(x^{(i)}\right) = y^{(i)}$ for all $i = 1, \dots, d$
  - Given a hypothesis class $\mathcal{H}$, define its ***Vapnik-Chervonenkis dimension***, VC$(\mathcal{H})$ to be the size of the largest set that is shattered by $\mathcal{H}$
    - If $\mathcal{H}$ can shatter arbitrarily large sets, then VC$(\mathcal{H}) = \infty$
    - under the definition of the VC dimension, in order to prove that VC$(\mathcal{H})$ is at least d, we only need to show that there's *at least* one set of size d that $\mathcal{H}$ can shatter
- the following theorem, due to Vapnik, can then be shown
  - arguably the most important theorem in all of learning theory
  - Let $\mathcal{H}$ be given
  - let $d =$ VC$(\mathcal{H})$
  - then, with probability at least $1 - \delta$, we have that for all $h \in \mathcal{H}$,

$$\left|\mathcal{E}(h) - \hat{\mathcal{E}}(h)\right| \leq O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

  - thus, with probability at least $1 - \delta$, we also have that:

$$\mathcal{E}\left(\hat{h}\right) \leq \mathcal{E}\left(h^*\right) + O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

  - i.e., if a hypothesis class has finite VC dimension, then uniform convergence occurs as m becomes large
    - as for the finite case, this allows us to give a bound on $\mathcal{E}(h)$ in terms of $\mathcal{E}\left(h^*\right)$
  - Corollary: For $\left|\mathcal{E}(h) - \hat{\mathcal{E}}(h)\right| \leq \gamma$ to hold for all $h \in \mathcal{H}$ (and hence $\mathcal{E}\left(\hat{h}\right) \leq \mathcal{E}(h^*) + 2\gamma$) with probability at least $1 - \delta$, it suffices that $m = O_{\gamma,\delta}(d)$.
    - i.e., the number of training examples needed to learn "well" using $\mathcal{H}$ is linear in the VC dimension of $\mathcal{H}$
      - for "most" hypothesis classes, the VC dimension (assuming a "reasonable" parameterization) is also roughly linear in the number of parameters
    - conclusion: for an algorithm that tries to minimize training error, the number of training examples needed is usually roughly linear in the number of parameters of $\mathcal{H}$