Recitation 9
Learning Theory, MLE/MAP, PCA

10-301/10-601: Introduction to Machine Learning

05/05/2021

1 Learning Theory

Some Important Definitions and Theorems

1. Basic notations:
   - **True function** (expert/oracle) $c^* : X \to Y$ (unknown)
   - **Hypothesis space** $\mathcal{H}$ and **hypothesis** $h \in \mathcal{H} : X \to Y$
   - Probability Distribution p^* (unknown)
   - Training Dataset $S = \{x^{(1)}, ...x^{(N)}\}$
2. **True Error (expected risk)**

$$R(h) = P_{x \sim p^*(x)}(c^*(x) \neq h(x))$$

3. **Train Error (empirical risk)**

$$\begin{aligned}\hat{R}(h) &= P_{x \sim S}(c^*(x) \neq h(x)) \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(c^*(x^{(i)}) \neq h(x^{(i)})) \\ &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(y^{(i)}) \neq h(x^{(i)}))\end{aligned}$$

4. **PAC criterion** is that we produce a high accuracy hypothesis with high probability. More formally,

$$P(\forall h \in \mathcal{H}, |R(h) - \hat{R}(h)| \leq \epsilon) \geq 1 - \delta$$

5. A hypothesis $h \in \mathcal{H}$ is **consistent** with training data S if $\hat{R}(h) = 0$ (zero training error/correctly classify)
6. **Sample Complexity** is the minimum number of training examples N such that PAC criterion is satisfied for a given ϵ (arbitrarily small error) and δ (with high probability)
7. Sample Complexity for 4 Cases: See Figure 1. Note that

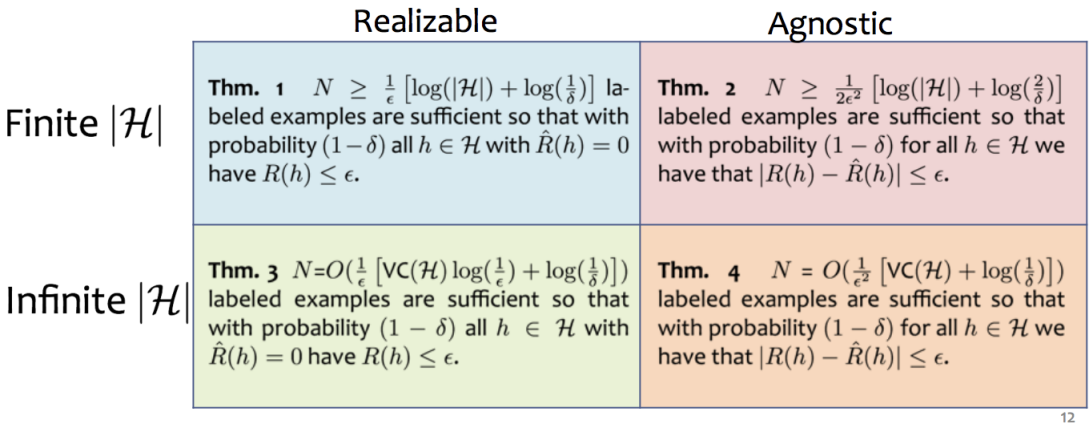


Figure 1

- **Realizable** means $c^* \in \mathcal{H}$
- **Agnostic** means c^* may or may not be in $\mathcal{H}$

8. **VC dimension** of a hypothesis space $\mathcal{H}$ is the maximum number of points such that there exists at least one arrangement of these points and a hypothesis $h \in \mathcal{H}$ that is consistent with any labelling of this arrangement of points.
9. If $\text{VC}(\mathcal{H}) = n$, then for all placements of $n+1$ points, there exists no hypothesis $h \in \mathcal{H}$ that can shatter any of it.

Questions

1. For the following examples, write whether or not the data points can be shattered using a linear classifier
   - 2 points in 1D
   - 3 points in 1D
   - 3 points in 2D
   - 4 points in 2D

   How many points can a linear boundary (with bias) classify exactly for d-Dimensions?

   - Yes
   - No
   - Yes
   - No

   $d+1$

2. Consider a semicircle classifier, where all points within the semicircle must equal 1 and all points outside must equal -1 (Rotated and scaled semi-circles are valid)

   (a) Which of the configurations of 4 points in figure 2 can a semicircle shatter?

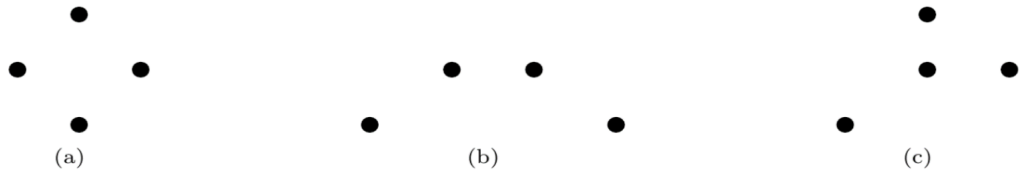


Figure 2

(a), (b)

   (b) What about the configurations of 5 points in figure 3?

Figure 3

None of the above (easy to construct counter examples)

3. Let $x_1, x_2, ..., x_n$ be n random variables that represent binary literals ($x \in \{0,1\}^n$). Let the hypothesis class H_n denote the conjunctions of no more than n literals in which each variable occurs at most once.

   Example: For $n = 4$, $x_1 \wedge x_2 \wedge x_4 \in H_4$

   Find the minimum number of examples required to learn $h \in H_{10}$ which guarantees at least 99% accuracy with at least 98% confidence.

   $|H_n| = 3^n$

   $|H_{10}| = 3^{10}, \epsilon = 0.01, \delta = 0.02$

   $N(H_{10}, \epsilon, \delta) \geq \lceil \frac{1}{\epsilon}[\ln |H_{10}| + \ln \frac{1}{\delta}] \rceil = \lceil 1489.81 \rceil = 1490$

2 MLE/MAP

As a reminder, in MLE, we have

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max_{\theta} p(\mathcal{D}|\theta) \\ &= \arg\min_{\theta} -\log\left(p(\mathcal{D}|\theta)\right)\end{aligned}$$

For MAP, we have

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|\mathcal{D}) \\ &= \arg\max_{\theta} \frac{p(\mathcal{D}|\theta)p(\theta)}{\text{Normalizing Constant}} \\ &= \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta) \\ &= \arg\min_{\theta} -\log\left(p(\mathcal{D}|\theta)p(\theta)\right)\end{aligned}$$

Imagine you are a data scientist working for an advertising company. The advertising company has recently run an ad and they want you to estimate it's performance. The ad was shown to N people. $Y^{(i)} = 1$ if person i clicked on the ad and 0 otherwise. Thus $\sum_i^N y^{(i)} = k$ people decided to click on the ad. Assume that the probability that the ith person clicks on the ad is θ and the probability that the ith person does not click on the ad is $1 - \theta$.

1. Note
$$p(\mathcal{D}|\theta) = p((Y^{(1)}, Y^{(2)}, ..., Y^{(N)}|\theta) = \theta^k(1-\theta)^{N-k}$$
Calculate $\hat{\theta}_{MLE}$.
$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\min_{\theta} -\log\left(p(\mathcal{D}|\theta)\right) \\ &= \arg\min_{\theta} -\log\left(\theta^k(1-\theta)^{N-k}\right) \\ &= \arg\min_{\theta} -k * \log(\theta) - (N-k)\log(1-\theta)\end{aligned}$$
Setting the derivative equal to zero yields
$$\begin{aligned}0 &= \frac{-k}{\theta} + \frac{(N-k)}{1-\theta} \\ \implies \theta_{MLE} &= \frac{k}{N}\end{aligned}$$

2. Your coworker tells you that $\theta \sim \text{Beta}(\alpha, \beta)$. That is:
$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

. note $B(\alpha, \beta)$ is not a function of θ and can be treated as a constant. Calculate $\hat{\theta}_{MAP}$

$$\begin{aligned}
\hat{\theta}_{MAP} &= \arg\min_{\theta} -\log\left(p(\mathcal{D}|\theta)p(\theta)\right) \\
&= \arg\min_{\theta} -\log\left(\frac{\theta^k(1-\theta)^{N-k}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)})\right) \\
&= \arg\min_{\theta} -\log\left(\theta^k(1-\theta)^{N-k}\theta^{\alpha-1}(1-\theta)^{\beta-1})\right) \\
&= \arg\min_{\theta} -\log\left(\theta^{k+\alpha-1}(1-\theta)^{N-k+\beta-1})\right) \\
&= \arg\min_{\theta} -(k+\alpha-1)\log(\theta) - (N-k+\beta-1)\log(1-\theta)
\end{aligned}$$

Setting the derivative equal to zero yields

$$\begin{aligned}
0 &= \frac{-k-\alpha+1}{\theta} + \frac{(N-k+\beta-1)}{1-\theta} \\
\implies \theta_{MAP} &= \frac{k+\alpha-1}{N+\alpha+\beta-2}
\end{aligned}$$

3. Suppose N = 100 and k = 10. calculate $\hat{\theta}_{MLE}$

   $\hat{\theta}_{MLE} = \frac{k}{N} = 0.10$

4. Suppose N = 100 and k = 10. Furthermore, you believe that in general people click on ads about 6 percent of the time, so you, somewhat naively, decide to set $\alpha = 6+1 = 7$, and $\beta = 100 - 6 + 1 = 95$. calculate $\hat{\theta}_{MAP}$

   $\hat{\theta}_{MAP} = \frac{k+\alpha-1}{N+\alpha+\beta-2} = \frac{10+7-1}{100+102-2} = \frac{16}{200} = 0.08$

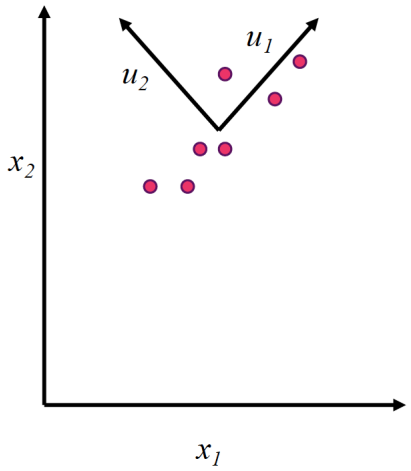
5. How do $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MAP}$ differ? Argue which estimate you think is better.

   Both estimates are reasonable given the available information. Note that $\hat{\theta}_{MAP}$ has lower variance than $\hat{\theta}_{MLE}$, but $\hat{\theta}_{MAP}$ is more biased. If you believe that this advertisement is similar to those advertisements that averaged a 6 percent click rate, then $\hat{\theta}_{MAP}$ may be a superior estimate, but if the circumstances under which the advertisement was shown were different from the usual, then $\hat{\theta}_{MLE}$ might be a better choice.

3 PCA

Principal Component Analysis aims to project data into a lower dimension, while preserving as much as information as possible.

How do we do this? By finding an orthogonal projection to the data that minimizes the squared error in reconstructing original data. In other words, find a new coordinate system.



Finding the vectors is quite easy visually as seen above, but how do we do this mathematically? We find the orthogonal vectors $u_1 \dots u_M$ such that it minimizes the sum of the errors of $||x^n - \hat{x}^n||^2$, where x^n is the given data point and $\hat{x}^n$ is the new vector (new dimension).

PCA: given M<d. Find $\langle \mathbf{u}_1 \dots \mathbf{u}_M \rangle$

that minimizes $E_M \equiv \sum_{n=1}^{N} ||\mathbf{x}^n - \hat{\mathbf{x}}^n||^2$

where $\hat{\mathbf{x}}^n = \bar{\mathbf{x}} + \sum_{i=1}^{M} z_i^n \mathbf{u}_i$

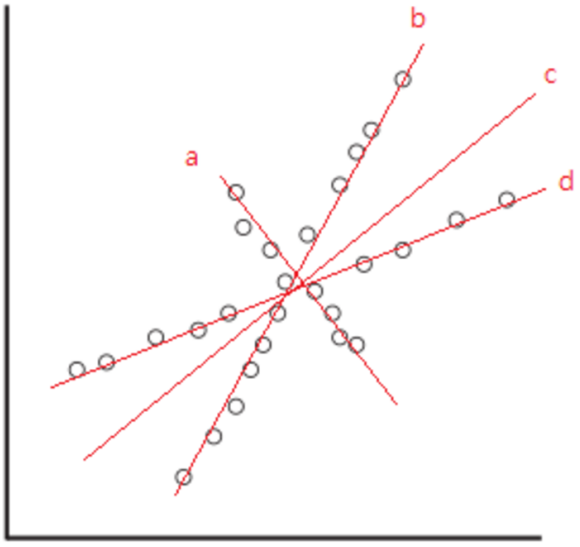
Mean

$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^n$

If we have M new vectors, and d original vectors, where $M < d$, it is not possible to reconstruct the original data without losing any error. In other words, if $M = d$, we can reconstruct the original data with 0 error. So, we know that all the error comes from the

(M-d) missing components (vectors). This error can be expressed in terms of the covariance matrix of the original data, and the error is minimized when each vector $u_1 \dots u_M$ are eigenvectors of the covariance matrix (slide 9 from lecture 27). The higher the eigenvalues are, the more information it stores (less error).

Use the figure below to answer the following questions.



1. What will be its first principal component? **Select one:**

   ◯ a

   ◯ b

   ◯ c

   ◯ d

   c

2. **NOTE : This is continued from the previous question.** What is the second principal component in the figure from the previous question? **Select one:**

   ◯ a

   ◯ b

   ◯ c

   ◯ d

   a

3. **NOTE : This is continued from the previous question.** What is the third principal component in the figure from the previous question? **Select one:**

   ◯ a

   ◯ b

   ◯ c

   ◯ d

   ◯ None of the above

   E : None of the above

4. Assume we are given a dataset X for which the eigenvalues of the covariance matrix are: (2.2, 1.7, 1.4, 0.8, 0.4, 0.2, 0.15, 0.02, 0.001). What is the smallest value of k we can use if we want to retain 75% of the variance (sum of all the variances in value) using the first k principal components?

   [ ]

   3

5. Assume we apply PCA to a matrix $X \in R^{n \times m}$ and obtain a set of PCA features, $Z \in R^{n \times m}$. We divide this set into two, $Z1$ and $Z2$. The first set, Z1, corresponds to the top principal components. The second set, Z2, corresponds to the remaining principal components. Which is more common in the training data: **Select one:**

   ◯ a point with large feature values in $Z1$ and small feature values in $Z2$

   ◯ a point with large feature values in $Z2$ and small feature values in $Z1$

   ◯ a point with large feature values in $Z2$ and large feature values in $Z1$

   ◯ a point with small feature values in $Z2$ and small feature values in $Z1$

   A