Recitation 6
Probabilistic Learning, CNNs, Learning Theory

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

11/02/2022

1 Probabilistic Learning

In probabilistic learning, we are trying to learn a target probability distribution as opposed to a target function. We'll review two ways of estimating the parameters of a probability distribution, as well as one family of probabilistic models: Naive Bayes classifiers.

1.1 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

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

---

1. Imagine you are a data scientist working for an advertising company. The advertising company has recently run an ad and wants you to estimate its performance.

   The ad was shown to N people. Let $Y^{(i)} = 1$ 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 i-th person clicks on the ad is θ and the probability that the i-th person does not click on the ad is $1 - \theta$.

(a) Note that

$$p(\mathcal{D}|\theta) = p((Y^{(1)}, Y^{(2)}, ..., Y^{(N)}|\theta) = \theta^k(1-\theta)^{N-k}$$

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

$$\begin{aligned}0 &= \frac{-k}{\theta} + \frac{(N-K)}{1-\theta} \\ \implies \hat{\theta}_{MLE} &= \frac{k}{N}\end{aligned}$$

(b) Suppose $N = 100$ and $k = 10$. Calculate $\hat{\theta}_{MLE}$.

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

(c) Your coworker tells you that $\theta \sim \text{Beta}(\alpha, \beta)$. That is:

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

Recall from lecture that $\hat{\theta}_{MAP}$ for a Bernoulli random variable with a Beta prior is given by:

$$\hat{\theta}_{MAP} = \frac{k + \alpha - 1}{N + \alpha + \beta - 2}$$

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 $\hat{\theta}_{MAP}$.

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

(d) How do $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MAP}$ differ in this scenario? 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 advertisements with 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.

2. Suppose you are an avid Neural and Markov fan who monitors the @neuralthenarwhal Instagram account each day. Suppose you wish to find the probability that Neural or Markov will post at any time of day. Over three days you look on Instagram and find the following number of new posts: $x = [3, 4, 1]$

A fellow fan tells you that this comes from a Poisson distribution:

$$p(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}$$

Also, you are told that $\theta \sim \text{Gamma}(2, 2)$ — that is, its pdf is:

$$p(\theta) = \frac{1}{4}\theta e^{-\frac{\theta}{2}},\ \theta > 0$$

Calculate $\hat{\theta}_{MAP}$.

(See also https://en.wikipedia.org/wiki/Conjugate_prior)

Note:

$$p(\mathcal{D}|\theta) = \frac{e^{-\theta}\theta^3}{3!}\frac{e^{-\theta}\theta^4}{4!}\frac{e^{-\theta}\theta^1}{1!}$$

$$\begin{aligned}
\hat{\theta}_{MAP} &= \arg\min_\theta - \log\left(p(\mathcal{D}|\theta)p(\theta)\right) \\
&= \arg\min_\theta - \log\left(\frac{e^{-\theta}\theta^3}{3!}\frac{e^{-\theta}\theta^4}{4!}\frac{e^{-\theta}\theta^1}{1!} \times \frac{1}{4}\theta e^{-\frac{\theta}{2}}\right) \\
&= \arg\min_\theta - \log\left(\frac{e^{-3\theta-\frac{\theta}{2}}\theta^9}{3! \times 4!}\right) \\
&= \arg\min_\theta - \left(\left(-3\theta - \frac{\theta}{2}\right)\log e + 9\log\theta - \log(3! \times 4!)\right) \\
&= \arg\min_\theta \left(3\theta + \frac{\theta}{2}\right) - 9\log\theta + \log(3! \times 4!)
\end{aligned}$$

Taking the derivative gives us

$$\frac{d}{d\theta}\left(3\theta + \frac{\theta}{2}\right) - 9\log\theta + \log(3! \times 4!) = \left(3 + \frac{1}{2}\right) - \frac{9}{\theta}$$

Setting the derivative equal to zero yields

$$0 = \left(3 + \frac{1}{2}\right) - \frac{9}{\theta}$$

$$\implies \theta_{MAP} = \frac{9}{3 + \frac{1}{2}} = 2.57142857143$$

1.2 Naive Bayes

By applying Bayes' rule, we can model the probability distribution $P(Y|X)$ by estimating $P(X|Y)$ and $P(Y)$.

$$P(Y|X) \propto P(Y)P(X|Y)$$

The Naive Bayes assumption greatly simplifies estimation of $P(X|Y)$ - we assume the features X_d are independent given the label. With math:

$$P(X|Y) = \underline{\hspace{6cm}}$$

Different Naive Bayes classifiers are used depending on the type of features.

- Binary Features: Bernoulli Naive Bayes - $X_d \,|\, Y = y \sim \text{Bernoulli}(\theta_{d,y})$
- Discrete Features: Multinomial Naive Bayes - $X_d \,|\, Y = y \sim \text{Multinomial}(\theta_{d,1,y}, \ldots, \theta_{d,K-1,y})$
- Continuous Features: Gaussian Naive Bayes - $X_d \,|\, Y = y \sim \mathcal{N}(\mu_{d,y}, \sigma^2_{d,y})$

We'll walk through the process of learning a Bernoulli Naive Bayes classifier. Consider the dataset below. You are looking to buy a car; the label is 1 if you are interested in the car and 0 if you aren't. There are three features: whether the car is red (your favorite color), whether the car is affordable, and whether the car is fuel-efficient.

Interested?	Red?	Affordable?	Fuel-Efficient?
1	1	1	1
0	0	1	0
0	0	1	1
1	0	0	0
0	0	1	1
0	0	1	1
1	1	1	1
1	1	0	1
0	0	0	0

1. How many parameters do we need to learn?

   6 for $P(X|Y)$, 1 for $P(Y)$

2. Estimate the parameters via MLE.

	$Y = 1$	$Y = 0$
Red?	$\frac{3}{4}$	0
Affordable?	$\frac{1}{2}$	$\frac{4}{5}$
Fuel-Efficient?	$\frac{3}{4}$	$\frac{3}{5}$

3. If I see a car that is red, not affordable, and fuel-efficient, would the classifier predict that I would be interested in it?

   $P(Y = 1|\text{red, not affordable, efficient}) \propto \frac{4}{9} \cdot \frac{3}{4} \cdot \frac{2}{4} \cdot \frac{3}{4} = \frac{1}{8}$

   $P(Y = 0|\text{red, not affordable, efficient}) \propto \frac{5}{9} \cdot 0 \cdot \frac{1}{5} \cdot \frac{3}{5} = 0$

4. Is there a problem with this classifier based on your calculations for the previous question? If so, how can we fix it?

   If the car is red, the classifier will always predict I'm interested because $P(\text{not red}|Y = 0) = 0$. We can use a prior which prevents parameter estimates from being 0, i.e. adding 1 fake count for each feature/label combination. This will be important in Homework 7!

5. Now we will derive the decision boundary of a 2D Gaussian Naïve Bayes. Show that this decision boundary is quadratic. That is, show that $p(y = 1 \mid x_1, x_2) = p(y = 0 \mid x_1, x_2)$ can be written as a polynomial function of x_1 and x_2 where the degree of each variable is at most 2. You may fold *unimportant* constants into terms such as C, C', C'', C''' so long as *you are clearly showing each step.*

   Observe that both the LHS and RHS should equal $\frac{1}{2}$ at the decision boundary, so they are both nonzero.

$$
\begin{aligned}
&p(y = 1 \mid x_1, x_2) = p(y = 0 \mid x_1, x_2) \\
\implies &\frac{p(x_1 \mid y = 0)p(x_2 \mid y = 0)p(y = 0)}{p(x_1, x_2)} = \frac{p(x_1 \mid y = 1)p(x_2 \mid y = 1)p(y = 1)}{p(x_1, x_2)} \\
\implies &1 = \frac{p(x_1 \mid y = 1)p(x_2 \mid y = 1)p(y = 1)}{p(x_1 \mid y = 0)p(x_2 \mid y = 0)p(y = 0)} \qquad (\because \text{nonzero LHS}) \\
\implies &1 = C \exp\left[\frac{(x_1 - \mu_{11})^2}{2\sigma_{11}^2} + \frac{(x_2 - \mu_{21})^2}{2\sigma_{21}^2} - \frac{(x_1 - \mu_{10})^2}{2\sigma_{10}^2} - \frac{(x_2 - \mu_{20})^2}{2\sigma_{20}^2}\right] \\
\implies &0 = C' + \frac{(x_1 - \mu_{11})^2}{2\sigma_{11}^2} + \frac{(x_2 - \mu_{21})^2}{2\sigma_{21}^2} - \frac{(x_1 - \mu_{10})^2}{2\sigma_{10}^2} - \frac{(x_2 - \mu_{20})^2}{2\sigma_{20}^2} \qquad (\because \text{nonzero } C)
\end{aligned}
$$

   Since C' is some constant that does not depend on x_1 or x_2, we have shown that the decision boundary is (at most) quadratic x_1 and x_2.

2 Learning Theory

2.1 PAC Learning

Some Important Definitions

1. Basic notation:
   - Probability distribution (unknown): $X \sim p^*$
   - **True function** (unknown): $c^* : X \to Y$
   - **Hypothesis space** $\mathcal{H}$ and **hypothesis** $h \in \mathcal{H} : X \to Y$
   - Training dataset $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$
2. **True Error (expected risk)**

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

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

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

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

$$P(\forall h \in \mathcal{H}, \underline{\hspace{6em}} \leq \underline{\hspace{2em}}) \geq \underline{\hspace{4em}}$$

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

Sample Complexity is the minimum number of training examples N such that the PAC criterion is satisfied for a given ϵ and δ

Sample Complexity for 4 Cases: See Figure 1. Note that

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

	Realizable	Agnostic
Finite $\|\mathcal{H}\|$	**Thm. 1** $N \geq \frac{1}{\epsilon}\left[\log(\|\mathcal{H}\|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 2** $N \geq \frac{1}{2\epsilon^2}\left[\log(\|\mathcal{H}\|) + \log(\frac{2}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\|R(h) - \hat{R}(h)\| \leq \epsilon$.
Infinite $\|\mathcal{H}\|$	**Thm. 3** $N = O(\frac{1}{\epsilon}\left[\mathrm{VC}(\mathcal{H})\log(\frac{1}{\epsilon}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \epsilon$.	**Thm. 4** $N = O(\frac{1}{\epsilon^2}\left[\mathrm{VC}(\mathcal{H}) + \log(\frac{1}{\delta})\right])$ labeled examples are sufficient so that with probability $(1-\delta)$ for all $h \in \mathcal{H}$ we have that $\|R(h) - \hat{R}(h)\| \leq \epsilon$.

12

Figure 1: Sample Complexity for 4 Cases

The **VC dimension** of a hypothesis space $\mathcal{H}$, denoted $\mathrm{VC}(\mathcal{H})$ or $d_{VC}(\mathcal{H})$, 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.

To show that $\mathrm{VC}(\mathcal{H}) = n$:

- Show there exists a set of points of size n that $\mathcal{H}$ can shatter
- Show $\mathcal{H}$ cannot shatter any set of points of size $n+1$

Questions

1. For the following examples, write whether or not there exists a dataset with the given properties that can be shattered by 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 rectangle classifier (i.e. the classifier is uniquely defined 3 points $x_1, x_2, x_3 \in \mathbb{R}^2$ that specify 3 out of the four corners), where all points within the rectangle must equal 1 and all points outside must equal -1

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

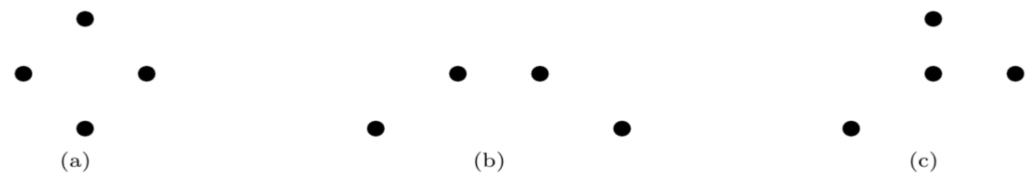

Figure 2

(a), (b), since the rectangle can be scaled and rotated it can always perfectly classify the points. (c) is not perfectly classifiable in the case that all the exterior points are positive and the interior point is negative.

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

Figure 3

None of the above. For (d), consider (from left to right) the labeling 1, 1 -1, -1, 1. For (e), same issue as (c).

3. In the below table, state in which case the sample complexity of the hypothesis falls under.

<table>
<tr><th>Problem</th><th>Hypothesis Space</th><th>Realizable/ Agnostic</th><th>Finite/ Infinite</th></tr>
<tr><td>A binary classification problem, where the data points are linearly separable</td><td>Set of all linear classifiers</td><td></td><td></td></tr>
<tr><td>Predict whether it will rain or not based on the following dataset:
<table>
<tr><th>Temp</th><th>Humid</th><th>Wind</th><th>Rain?</th></tr>
<tr><td>High</td><td>Yes</td><td>Yes</td><td>Yes</td></tr>
<tr><td>Low</td><td>Yes</td><td>No</td><td>No</td></tr>
<tr><td>Low</td><td>No</td><td>Yes</td><td>Yes</td></tr>
<tr><td>High</td><td>No</td><td>No</td><td>Yes</td></tr>
</table></td><td>A decision tree with max depth 2, where each node can only split on one feature, and the features cannot be repeated along a branch</td><td></td><td></td></tr>
<tr><td>Classifying a set of real-valued points where the underlying data distribution is unknown</td><td>Set of all linear classifiers</td><td></td><td></td></tr>
<tr><td>A binary classification problem on a given set of data points, where the data is not linearly separable</td><td>K-nearest neighbour classifier with Euclidean distance as distance metric</td><td></td><td></td></tr>
</table>

	Realizable/ Agnostic	Finite/ Infinite
1	Realizable	Infinite (All possible linear classifiers)
2	Realizable (We can split the given data using a depth 2 decision tree)	Finite (There are only a finite set of decision trees that can be formed with the given constraints)
3	Agnostic (The data may or may not be linearly separable)	Infinite
4	Agnostic (The KNN classifier may or not be able to perfectly classify each point)	Finite (The hypothesis space is the set of all possible partitions of the input space into k-nearest regions - which is finite for all possible values of k)

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

   Example: For $n = 4$, $(x_1 \wedge x_2 \wedge x_4), (x_1 \wedge \neg x_3) \in \mathcal{H}_4$

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

   $|H_n| = 3^n$

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

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

3 Generative vs. Discriminative Models

1. What is the difference between discriminative vs. generative models?
   The discriminative model maximizes the conditional likelihood: $P(Y|X)$.
   The generative model maximizes the joint likelihood of the observations x and the labels y: $P(X, Y)$.

2. Classify logistic regression and Naive Bayes as discriminative vs. generative model.
   Logistic Regression directly estimates the parameters of $P(Y|X)$, whereas Naive Bayes directly estimates $P(X, Y)$ via parameters for $P(Y)$ and $P(X|Y)$. We often call the former a discriminative classifier, and the latter a generative classifier.

3. We say that two models form a generative/discriminative pair if the conditional distribution $P(Y|X)$ inferred from the joint $P(X, Y)$ of the generative model is equivalent to the $P(Y|X)$ of the discriminative. Gaussian Naive Bayes and Logistic Regression are one such pair.
   Describe the tradeoffs of generative vs. discriminative models.
   (Another pair we won't discuss are HMMs and linear-chain CRFs.)

   Assume that we learn a generative/discriminative pair of models from a finite training dataset. If model assumptions are correct, the generative model (e.g. Naive Bayes) is a more efficient learner and requires fewer samples than its discriminative counterpart (e.g. logistic regression). If model assumptions are incorrect, then the discriminative model (e.g. logistic regression) has lower asymptotic error (as the training size grows) and does better than its generative counterpart (e.g. Naive Bayes).

   An example of an easily violable assumption would be that the features x are conditionally independent given y (since that is rarely the case in reality).