COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Pranay Manocha

Lecture #19
April 15, 2019

1 Introduction and Motivation

For the first part of the course we focused on classification learning, until last week where we discussed regression problems and trying to estimate a real-valued number. Today we turn to look at the question of how to model a probability distribution. In the past, we have looked at problems where we assume a distribution (like in the PAC learning setting), but we never decided to find what the distribution is. In fact, we went further and assumed the results to be true over all distributions. In this new setting, we will get some samples x from a probability distribution. $x \sim P$ and then given the samples, the goal is to estimate P itself.

This model is of particular interest to statisticians and in general those who want to model, for instance, characteristics of a human population, such as the distribution of things like SAT scores, weights or heights of people and so on. It is also very useful in Natural Language Processing (NLP) to build models over the distribution of what is being said or written. Typically, as an example, we model the English sentences written by people as a probability distribution. If we know the model, we can see which of the given sentences are more probable than the others. This model can be especially useful in speech recognition where, if we model English utterances as coming from some complex probability distribution, and if we have some model of this distribution, then we can look for mistakes and correct them using this model. For example, consider these two transcriptions of a speech recognition system:

1. He sat on the chair.
2. He fat on the chair.

If we have a good model of the distribution of spoken sentences, then we can say that the first sentence is more probable than the second, and therefore is the correct transcription. These days, we are also seeing word suggestions and corrections while typing (google keyboard) which are very robust and accurate most of the time, so it is a very useful thing to build such a model.

Even for classification, these models can be useful. For instance, if we model the distributions of men and women heights, and then are given the height of a random person that should be classified as a man or women, we can predict the person is a man if and only if the probability of being a man, according to the modeled distributions, is more than the probability the person is a woman. This amounts to finding a threshold value, as in Fig. 1, such that the probability a person above that height should be labelled man is greater than .5 (assuming men are taller). Previously in the course, we would have just established such a threshold value directly (from using the given data, making as few mistakes on the data as possible), but here we would do so as a function of the distributions we compute.

The approach of modeling distributions, and thus how the data is generated, is more statistical in nature and is often called a generative approach, while the alternative of directly trying to find an accurate discriminator is said to be a discriminative approach. The discriminative approach is more direct, not trying to model the distribution which is not our goal, which is instead to be correct as often as possible. On the other hand, in the generative modelling approach, we might make some more assumptions about the data to get the right model, which may or may not be a bad thing. The advantage of this is that we generally need less data for training, because of the assumptions we make. If the assumptions we make are valid, we are able to get equally meaningful results using less data.

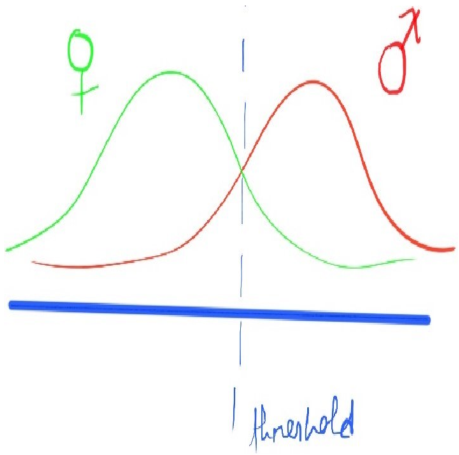

Figure 1: Probability distribution of men and women in example

2 Principle of Maximum Likelihood

So the problem we are trying to solve here is as follows: we have an unknown distribution P, and we also have m samples from that distribution sampled iid. Our aim now is to have a good model for P using the m samples that we have. For simplicity, we assume that the unknown distribution P is discrete (finite or countably infinite). This problem is called Density Estimation or Probability Modelling.

As in the PAC model, let's assume we have a set $\mathcal{Q}$ which is a set of distributions and we want to find the distribution in $\mathcal{Q}$ that is as close to P as possible. Analogously, we can imagine this to be similar to the hypothesis set we had earlier in PAC learning. Let's see an example here in the extreme case where we have two possible distributions that could be generating P.

2.1 Example with distributions

We have the distributions q_1 and q_2 given in Fig. 2. Based on these two distributions, we need to tell which of these two distributions are more likely to be the model that generated these points. The answer here is simple: we see here that the points are more aligned towards the distribution q_2 than q_1 and so it is more likely that the data model that could have generated these points is q_2. In a more general sense, we see that we can check each candidate hypothesis distribution q_i and find the probability of those points being generated if q_i were the actual distribution, and then select the distribution for which this probability is highest.

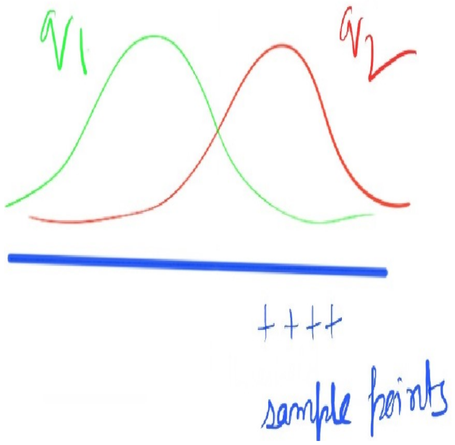

Figure 2: Sample Probability Distributions

Or more generally, if we suppose that q is the candidate distribution that generated the data, then

$$\begin{aligned} Pr_q[x_1, x_2...x_m] &= q(x_1)q(x_2)....q(x_m) \\ &= \prod_{i=1}^{m} q(x_i) \end{aligned} \tag{1}$$

where all points x_i are assumed to be iid. This probability is called the *likelihood* of the data under q. We then choose q which has the highest likelihood.

2.1.1 Distribution of Outcomes of Coin Flips

Let's take the example of finding the distribution of outcome of coin flips, which essentially means finding the bias of the coin. Each example x has two possible outcomes

$$X = \left\{ \begin{array}{cc} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{array} \right\}$$

The space of distributions $\mathcal{Q}$ is [0,1], because the bias can be anything between this range. We gather some data (m samples). Suppose we flip the coin m times and get heads in h of these flips. Another way of writing this is $h = \sum_{i=1}^{m} x_i$. Then if $P = q$ (true bias is equal to q, which means that we are assuming the data to be generated from q), we can compute the likelihood of the observed coin flips to be:

$$\prod_{i=1}^{m} q(x_i) = \prod_{i=1}^{m} \left\{ \begin{array}{ll} q, & \text{when } x_i = 1 \\ 1-q, & \text{when } x_i = 0 \end{array} \right\}$$

which becomes

$$q^h(1-q)^{m-h} \tag{2}$$

Our aim is to maximise this expression over q, to get q which maximizes Eq. (2). Hence, the problem becomes

$$\max_q \quad q^h(1-q)^{m-h} \tag{3}$$

We can solve this by differentiating this expression over q and equating to 0 to get the optimal maximum-likelihood value of q. Solving, we get

$$q = h/m \tag{4}$$

which is one of the obvious ways of estimating the bias, which says that the bias is equal to the fraction of the number of heads in the samples that we collected.

We now formally state what the Principle of Maximum Likelihood is.
We solve for:

$$\begin{aligned}
\max_{q\in\mathcal{Q}} \prod_{i=1}^{m} q(x_i) &\equiv \max_{q\in\mathcal{Q}} \log \prod_{i=1}^{m} q(x_i) \\
&\equiv \max_{q\in\mathcal{Q}} \sum_{i=1}^{m} [\log q(x_i)] \\
&\equiv \min_{q\in\mathcal{Q}} \sum_{i=1}^{m} [-\log q(x_i)] \\
&\equiv \min_{q\in\mathcal{Q}} \frac{1}{m} \sum_{i=1}^{m} [-\log q(x_i)]
\end{aligned}$$

Here we turned the problem of Maximum likelihood to a problem of minimization over the seen samples. We have seen this kind of problem before. In classification, we were minimizing the training error, which is the average error over all seen samples (zero-one loss). When we were doing regression, we were minimizing the average of square loss. Intuitively, loss depicts some kind of discrepancy between the model and the sample point.

We see that here, $-\log q(x_i)$ is a measure of how poorly q fits x_i, i.e., the discrepancy between the model and data; we will call it the log loss function. This log loss function is an especially important loss function. Here we will thus have the average of log loss over the samples, which is the empirical risk for the log loss function of q. We also note that the empirical risk should give some close estimate of the true expected loss.

When the true distribution is P, we can compute the true risk for model q under the log loss to be:

$$\begin{aligned}
TrueRisk &= \mathbb{E}_{x\sim P}[-\log q(x)] \\
&= -\sum_{x\in\mathcal{X}} [P(x) \log q(x)] \qquad (5)
\end{aligned}$$

where $\mathcal{X}$ denotes the entire space, which we here assume is finite (or at least countable) for simplicity. This entire term is sometimes referred to as cross entropy.

$$\begin{aligned}
&= \sum_{x\in\mathcal{X}} \left[P(x) \log\left(\frac{P(x)}{q(x)}\right) - P(x)\log P(x)\right] \\
&= RE(P||q) + H(P) \qquad (6)
\end{aligned}$$

where RE stands for relative entropy between P and q and $H(P)$ is the entropy of P. Hence, the true risk (using log loss) can be written as the sum of relative entropy between

P and q and the entropy of P. Note that $H(P)$ does not depend on q, so minimizing true risk is equivalent to minimizing $RE(P||q)$ over $\mathcal{Q}$ which means that we are trying to find q as close as possible to P, as relative entropy is always non-negative and is zero if and only if both the distributions are the same.

3 Maximum Entropy Modelling of Distributions

We now consider a more practical setting. Consider the problem of modeling the habitat of plant/animal species. Perhaps you are a researcher on an island who has a sample of butterfly sightings, along with features associated with each sighting (for instance: altitude, annual rainfall, average temperature, etc.), and you wish to model the population distribution of the butterfly on the island. We make several assumptions: that there exists a true probability distribution D that would properly model the species, that the sightings are being sampled from this same distribution D, and that it is possible to get every bit of data for each feature for each spot on the map (our domain, X, although we first generally divide the map into a grid of cells, so that X is finite as in our above assumption). More formally, let $|X| = N$ and consider $x_1, ..., x_m \sim D$, and features $f_1, .., f_n$ such that these features are real-valued functions $f_j : X \to \mathbb{R}$, where our goal is to estimate the true distribution D. We begin by considering two different approaches. In the end, we show that both the approaches have the same solution.

3.1 Principle of Maximum Entropy

It is difficult to estimate the true distribution, so the easiest step is to begin with taking the average expectation of each of the features as an estimate of the true expectation for the features.

The expected value of f_j is

$$\mathbb{E}_D[f_j] = \mathbb{E}_{x\sim D}[f_j(x)] \tag{7}$$

whereas the empirical estimate from the samples is

$$\hat{\mathbb{E}}[f_j] = \frac{1}{m}\sum_{i=1}^{m} f_j(x_i) \tag{8}$$

Here D is unknown and we seek to find a distribution p that will be a good estimate of D. We expect $\hat{\mathbb{E}}[f_j]$ to be close to $\mathbb{E}_D[f_j]$ $\forall$ j. So therefore, it makes sense that the same should hold for our estimate p of D. This suggests that we should choose p to have the property that $\mathbb{E}_p[f_j]$, the expectation of f_j under the estimated distribution p, should be close to $\hat{\mathbb{E}}[f_j]$. We will go further and require equality here. That is, we will seek to find a distribution p for which $\mathbb{E}_p[f_j] = \hat{\mathbb{E}}[f_j]$ $\forall j \in \{1, \ldots, n\}$.

In the above expression, the left half of the equality denotes the expectation of feature f_j under distribution p and the right half of the equality denotes the empirical average that we see from the samples we observe. In terms of our example, for instance, if we only have found the butterfly at high altitudes, we find a p which predicts the same. We can rewrite our constraints on p to say that we require that p belong to the set $\mathcal{P}$ where

$$\mathcal{P} = \{p \mid \mathbb{E}_p[f_j] = \hat{\mathbb{E}}[f_j] \;\; \forall j\}$$

Another point to note is that the number of distributions that will satisfy these properties mentioned could be large and somehow we need to select just one. Given no prior beliefs and no observations, we would just guess that the most intuitive guess would be the uniform distribution, so maybe it would make sense to choose the distribution which is closest to the uniform distribution, among all distributions which satisfy the above constraints of $\mathcal{P}$. (If we have reason to choose some other distribution as the default, then the method we are describing can be generalized for that case as well, but we will only use uniform distribution for this purpose.) Hence, among all distributions p which satisfy:

$$\mathbb{E}_p[f_j] = \hat{\mathbb{E}}[f_j] \quad \forall j \tag{9}$$

we seek to find the one which is closest to uniform, that is, which minimizes $RE(p||\mathit{unif})$, where *unif* denotes the uniform distribution over X.

The relative entropy can be written down as:

$$\sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{1/N} \right) \tag{10}$$

which can be reduced to

$$\log N + \sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{11}$$

The second term here is the negative of the entropy, which becomes:

$$\log N - H(p) \tag{12}$$

Now the minimization problem turns into a maximization problem, namely

$$\arg\max_{p \in \mathcal{P}} H(p) \tag{13}$$

where $\mathcal{P} = \{\text{p} \mid \mathbb{E}_p[f_j] = \hat{\mathbb{E}}[f_j] \ \forall \text{j}\}$

In general, we can do better by combining the features (higher order features) and do better by creating a new set of features, for instance, by taking products or squares of features which is generally helpful in such problems.

One more interesting thing to note is that in this setting, we are not considering any proximity between the points as a factor in our decision, though this could help. Biologists though have a different view — they want to individually validate findings separately. We can add these extra feature vectors f which would provide location information where the butterfly was found as useful features into our model if we want.

3.2 Using Gibbs Distribution / Exponential Family Distribution

In the previous subsection, we said that we needed to find $p \in \mathcal{P}$ which maximizes $H(p)$. We did not say anything about the structure of the distribution. Here we instead assume that distribution we are looking for has a particular form. Perhaps it would be reasonable to assume that it is linear in each feature, except then we will end up with ill-defined probabilities (such as negative values). Thus, we instead use a linear function in the exponent, and re-scale accordingly. In this case, we use

$$q(x) = \exp\left(\sum_{j=1}^{n} \lambda_j f_j(x) \right) / Z_\lambda \tag{14}$$

for some setting of the real-valued parameters $\langle \lambda_1, \ldots, \lambda_n \rangle$ where, as stated, we use an exponential to avoid negative values, and we normalize with Z_λ to make it a probability distribution. This type of distribution is referred to as a Gibbs distribution or exponential-family distribution.

Let $\mathcal{Q}$ be all the distributions that have the above form. Now, we will use the principle of maximum likelihood: we want to find q such that it maximizes the log loss. Hence, $\arg\max_{q \in \mathcal{Q}} \sum_{i=1}^{m} \log[q(x_i)]$.

We quickly note that such a maximum may not be realised by a distribution q in $\mathcal{Q}$, and therefore, the maximum might not exist. Therefore, we also allow the distribution q to be in the closure of $\mathcal{Q}$, written $\overline{\mathcal{Q}}$, which includes all limits of sequences in $\mathcal{Q}$. (For instance, the set $A = [0, 1)$ has no maximum, but if we take its closure $\overline{A} = [0, 1]$, it now has a maximum, exactly equal to 1). So the problem becomes $\arg\max_{q \in \overline{\mathcal{Q}}} \sum_{i=1}^{m} \log[q(x_i)]$. (We also note here that *sup* does not work as we have to do arg max and not max. In the latter case, when we do not have to extract the value out of the set, sup would have worked as well.)

4 Equivalence and Uniqueness Results

In previous sections, we discussed two methods. We will summarise them here and then discuss the implications of these results.

Theorem 1 *The following are equivalent:*

1. $q^* = \arg\max_{p \in \mathcal{P}} H(p)$
2. $q^* = \arg\max_{q \in \overline{\mathcal{Q}}} \sum_{i=1}^{m} \log q(x_i)$
3. $q^* \in \mathcal{P} \cap \overline{\mathcal{Q}}$

Furthermore, any of these determine q^* uniquely.

We won't prove these results in detail. We will just give a sketch to provide intuition below.

We notice that condition 1 and condition 2 imply that if q^* is the maximum entropy solution subject to the constraints mentioned in condition 1, then the very same q^* distribution is also the maximum likelihood solution amongst all Gibbs distributions and also vice versa.

We see that condition 3 says that it is both a necessary and sufficient condition to find an element in the intersection of $\mathcal{P}$ and $\overline{\mathcal{Q}}$, and that this element is unique and is always a solution to both. The equivalence of the two approaches comes from them being duals of each other, as we will see the solution sketch below.

4.1 Sketch of Equivalence of Condition 1 and Condition 2

We see that condition 1 and condition 2 are duals of each other. We talked about duality in support vector machines between optimization problems. To see how, we use Lagrange multipliers to convert one optimization problem into the other.

We will start with condition 1, which we assume is our primal problem which is $q^* = \arg\max_{p \in \mathcal{P}} H(p)$. The Lagrangian is formed as follows:

$$L = \sum_{x \in \mathcal{X}} q(x) \log q(x) + \sum_{j=1}^{n} \lambda_j \left(\hat{\mathbb{E}}(f_j) - \sum_{x \in \mathcal{X}} q(x) f_j(x) \right) + \gamma \left(\sum_{x \in \mathcal{X}} q(x) - 1 \right) \tag{15}$$

The first term is the primal objective, the second term is used as:

$$\hat{\mathbb{E}}[f_j] - \mathbb{E}_q[f_j] = 0 \quad \forall j \tag{16}$$

which can be written as

$$\hat{\mathbb{E}}[f_j] - \sum_{x \in \mathcal{X}} q(x) f_j(x) = 0 \quad \forall j \tag{17}$$

In this setting, the λ_j's and γ are the Lagrange multipliers. The solution to this problem is a sadde point solution. As the problem now is to minimize L in terms of q but maximise L in terms of the λ_j's and γ. Taking derivative of L with respect $q(x)$, $\frac{dL}{dq(x)} = 0$, we get

$$1 + \log q(x) - \sum_{j=1}^{n} \lambda_j f_j(x) + \gamma = 0 \tag{18}$$

Solving this gives,

$$\begin{aligned} q(x) &= \exp\left(\sum_{j=1}^{n} \lambda_j f_j(x) - \gamma - 1 \right) && (19) \\ &= \exp\left(\sum_{j=1}^{n} \lambda_j f_j(x)) / \exp(\gamma + 1) \right) && (20) \\ &= \exp\left(\sum_{j=1}^{n} \lambda_j f_j(x) \right) / Z_\lambda && (21) \end{aligned}$$

where $\exp(\gamma + 1)$ acts as the normalization factor Z_λ. We see that this gives back the exponential family distribution and thus q has to be in the set $\mathcal{Q}$. We (partially) plug this value back into L and maximize with respect to the Lagrangian variables. We get:

$$\begin{aligned} L &= \sum_{x \in \mathcal{X}} q(x) \left(\sum_{j=1}^{n} \lambda_j f_j(x) - \log Z_\lambda \right) + \sum_{j=1}^{n} \lambda_j (\hat{\mathbb{E}}[f_j]) - \sum_{x \in \mathcal{X}} \left(q(x) \sum_{j=1}^{n} \lambda_j f_j(x) \right) && (22) \\ &= \frac{1}{m} \sum_j \left(\lambda_j \sum_i f_j(x_i) \right) - \log Z_\lambda && (23) \\ &= \frac{1}{m} \sum_i \left(\sum_j \lambda_j f_j(x_i) \right) - \log Z_\lambda && (24) \\ &= \frac{1}{m} \sum_i [\log q(x_i)] && (25) \end{aligned}$$

using Eq. (21). This is the log likelihood, or the negative empirical risk. Thus, at the solution, which is at a saddle point, the distribution will be a Gibbs distribution and it will have maximum likelihood/minimum log loss. This is exactly condition 2.