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 the question of how to model a probability distribution; this is called “density estimation.” In this model we will get some samples from a probability distribution $x \sim P$ and then given the samples, the goal is to estimate $P$ itself.

This has many advantages over our past distribution-free labelling-based model for statisticians and in general those who want to model things like scores on a standardized test. It can also be used for things like speech recognition, if we model English utterances as coming from some complex probability distribution, and if we have some model of this distribution, then we could program speech recognition software to make corrections using Bayes rule and its confidence that it heard a particular utterance, versus what it would expect in context. Of course, given a model of a distribution, we could use our model to predict labels. For instance, if we model the distributions of men/women heights, and then are given a random height that should be given one of the labels, we can find a threshold value 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, but here we would do so as a function of the distributions we compute.

2 The Principle of Maximum Likelihood

Given $x_1, x_2, ..., x_m \in P$ drawn iid from $P$ a discrete distribution, suppose we have a set of candidate distributions $Q$. For $q \in Q$, define $q(x) = \mathbb{P}_q[x]$. Then we define the "likelihood of the data under q" to be:

$$\mathbb{P}_q[x_1, ..., x_m] = q(x_1)q(x_2)...q(x_m) = \prod_{i=1}^{m} q(x_i)$$

We see that this “likelihood” quantifies how well $q$ fits $x_1, ..., x_m$ and thus we want to find the $q$ maximizing this product.

As a sanity check, we try a simple example. Consider flipping a coin $x$ that gives 1 with probability $p$ and 0 otherwise. We flip it $m$ times and get heads $h$ of these flips. Let $Q = \{ q/q is the bias of the coin \}$, then the likelihood under $q$ will be:

$$\prod_{i \in [m]} \{ q if x_i = 1 and (1 - q) otherwise \} = q^h(1 - q)^{m-h}$$
Thus to maximize the likelihood, we take the derivative with respect to $q$, set it equal to 0, giving $q = h/m$. This matches the intuition that our best guess for the bias of the coin is what we’ve observed.

In general, we want to choose $q \in Q$ with

$$\max_i \prod q(x_i) \equiv \max_i \log \prod q(x_i) \equiv \max_i \log q(x_i)$$

$$\equiv \min 1 \sum_{i=1}^m -\log q(x_i)$$

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

We define the “true risk” to be

$$E_{x \sim P}[-\log q(x)] = - \sum_{x \in P} P(x) \log q(x) = \sum_{x \in P} P(x) \log \frac{P(x)}{q(x)} - \sum_{x \in P} P(x) \log P(x)$$

$$= \text{RE}(P||q) + H(P)$$

where $H(P)$ is the entropy of $P$.

Note that $H(P)$ doesn’t depend on $q$, so minimizing true risk is equivalent to minimizing $\text{RE}(P||q)$ over $q \in Q$.

### 3 Maximum Entropy Modeling 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, 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 and that it’s 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, consider $x_1, \ldots, x_m \sim D$, and features $f_1, \ldots f_n$ such that $f_j : X \rightarrow \mathbb{R}$, where our goal is to estimate the true distribution $D$. We begin by considering two different approaches.

#### 3.1 Using the Principle of Maximum Entropy

Estimating the whole distribution is hard, so maybe we should begin by computing

$$\hat{E}[f_j] = \frac{1}{m} \sum_{i=1}^m f_j(x_i)$$
as an estimate for $\mathbb{E}_{x \sim D}[f_j(x)]$, the true expectation for the feature $f_j$. Thus we begin by finding $q$ such that $\mathbb{E}_q[f_j] = \hat{\mathbb{E}}[f_j]$ for each $f_j$. In terms of our example, for instance, if we only have found the butterfly at high altitudes, we find a $q$ which predicts the same.

This gives us a start, but now what? Given no prior beliefs, we would just guess 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 empirical conditions.

Thus we find $q$ such that for all $j$, $\mathbb{E}_q[f_j] = \hat{\mathbb{E}}[f_j]$ and minimize

$$\text{RE}(q||\text{unif}) = \sum_{x \in X} q(x) \log \frac{q(x)}{1/N} = \log N - H(q)$$

where $\text{unif}$ is the uniform distribution over $X$, and $H(q)$ is the entropy as before. We thus see that, since $\log N$ is a constant, this is equivalent to maximizing entropy. This is sometimes called the “principle of maximum entropy”, and it intuitively corresponds to being as spread out as possible.

We summarize this technique by saying we want to find $\arg\max_q (H(q))$ such that $q \in P$ where $P = \{q: \mathbb{E}_q[f_j] = \hat{\mathbb{E}}[f_j] \ \forall j\}$.

### 3.2 Using Exponential-Family/Gibbs Distributions

Another possible technique may be to assume that $q$, the distribution we’re looking for, has a particular form. Perhaps it would be reasonable to assume that it’s linear in each feature? In this case, we use:

$$q(x) = \frac{e^{\sum_{j=1}^n \lambda_j f_j(x)}}{Z_\lambda}$$

where we use an exponential to avoid negative values, and we normalize with $Z_\lambda$ to make it a probability distribution. Sometimes these are called “exponential family distributions” or “Gibbs distributions.” Let $Q$ be the distribution set $\{q\}$ such that they have the above form.

Now we will use the principle of maximum likelihood; thus we want to find the $q$ such that we have:

$$\max_{q \in Q} \sum_i \log q(x_i)$$

We must quickly note that technically, such a maximum may not exist, but in such a case there is a limit point of $Q$, i.e. we instead find the $q$ such that we have:

$$\text{Sup}_{q \in Q} \left(\sum_i \log q(x_i)\right) = \max_{q \in \bar{Q}} \sum_i \log q(x_i)$$

Where $\bar{Q}$ is the closure of $Q$, and Sup is the supremum.
3.3 Equivalence and Uniqueness Results

In fact, 3.1 and 3.2 have identical solutions, and furthermore, for $q^* = \max_{q \in \mathcal{P}} H(q)$ and $q^* = \arg\max_{q \in \mathcal{Q}} \sum_i \log q(x_i)$, then $q^* \in \mathcal{P} \cap \mathcal{Q}$ is unique! This is a very useful fact when trying to prove the convergence of algorithms in settings like this, as we will see in the next lecture. We will not prove this equivalence, but the equivalence of 3.1 and 3.2 comes from them being convex duals of each other, as we will sketch below.

Side note: It is difficult to know how good our solutions are in absolute terms, since this would be determined by $\text{RE}(\mathcal{P}||q) + H(P)$ and we don’t know $H(P)$, but it is easy to check how good solutions are relative to one another since subtracting our approximations of the first term cancels out the $H(p)$ and gives an accurate comparison.

3.3.1 Sketch of Equivalence of 3.1 and 3.2

We will do the sketch using Lagrange multipliers. We begin with the Lagrangian:

$$L = \sum_x q(x) \log q(x) + \sum_{j=1}^n \lambda_j \hat{E}[f_j] - \sum_x q(x)f_j(x) + \gamma(\sum_x q(x) - 1)$$

where the $q(x)$ are our primal variables, and the $\lambda_j$ and $\gamma$ are the dual variables (minimizing the difference between the empirical and true expectation of the features, and making the $q(x)$ sum to 1, respectively). First we minimize over the primal variables:

$$\frac{\partial L}{\partial q(x)} = 1 + \log q(x) - \sum_j \lambda_j f_j(x) + \gamma = 0$$

$$\implies q(x) = e^{\sum_j \lambda_j f_j(x) - \gamma - 1} = \frac{e^{\sum_j \lambda_j f_j(x)}}{Z_{\lambda}}$$

where $Z_{\lambda} = e^{\gamma + 1}$. Thus we see that it gives an exponential family distribution, and we plug this back into $L$ and then maximize with respect to the dual variables:

$$L = \sum_x q(x)(\sum_j \lambda_j f_j(x) - \log Z) - \sum_j \lambda_j \sum_x q(x)f_j(x) + \sum_j \lambda_j \hat{E}[f_j]$$

$$= -\log Z + \frac{1}{m} \sum_j \lambda_j \sum_i f_j(x) = \frac{1}{m} \sum_i (\sum_j \lambda_j f_j(x_i) - \log Z)$$

which will attain its maximum when the argument of the outer sum is

$$\sum_j \lambda_j f_j(x_i) - \log Z = \log q(x_i)$$

(at a saddle point), as in the second solution.