6.7900 Fall 2024: Lecture Notes 2

1 Empirical Risk Minimization

Recap: In the last class, we encountered the challenge that in most real-world scenarios we do not know the distribution of future data (X, Y). We hypothesized that we could use our training data to select the best h. However, in order to use the training data to estimate the best hypothesis, we need to make some assumptions about the relationship between training and future/test data.

The most common assumption made in machine learning is that all data points $(X^{(n)}, Y^{(n)})$ in both training and future data are sampled independently from the same underlying distribution (*I.I.D. samples*).

Note: Although predominant in machine learning applications, in almost every real-world problem this IID assumption is not true. For example, let's say we are predicting life expectancy. The distribution of life expectancy changes over time, so if one divides training and testing data by time the prediction will likely be off. However, as long as one is aware (and states) this assumption, the models we get out are typically still useful in real-world domains.

Empirical Distribution: $\hat{p}(x, y) = \frac{1}{N}\sum_{n=1}^{N} \delta_{x^{(n)}, y^{(n)}}(x, y)$. δ_k is a Dirac delta, a function which is zero at every point except from k and integrates to 1, and therefore it is used to formalize discrete probabilities over continuous space.

Empirical Risk Minimization: instead of minimizing the risk under the true future data distribution p, we choose h to minimize the *empirical risk* (i.e. the risk under the empirical distribution) over the training data:

$$\mathbb{E}_p[L(Y, h(X))] \approx \mathbb{E}_{\hat{p}}[L(Y, h(X))] = \frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)}))$$

Why might this be a good approximation? Thanks to the *law of large numbers*, as we get more and more data (formally, in the limit of $N \to \infty$) the empirical risk approaches the true risk.

Law of Large Numbers: Let $Z_1, Z_2, \ldots$ be IID random variables. Assume all necessary expectations exist. Then, with probability 1:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} Z_n = \mathbb{E}[Z_1]$$

See probability classes/textbooks for the difference between *surely* or *always* and *with probability 1* or *almost surely*.

However empirical risk minimization does not come without a new set of challenges as highlighted by the following example.

Example (overfitting): Take the spam detection example (label: 1 spam or 0 not-spam), 0-1 loss, and the following decision rule: $h(x) = 1$ if the timestamp of x matches the timestamp (assumed to be unique) of any spam email in the training set exactly, else 0.

By construction this decision rule will classify as spam all the spam emails in the training set and as not spam all the other emails, therefore the empirical risk of this decision rule on the training set is zero. Therefore, this will always be a decision rule minimizing the empirical risk. But is this a good decision rule? Given that it will classify as non-spam any email arriving in the future, this decision rule is clearly not useful. This leads us to refine, in general terms, our goal to that of *generalization* and to notice the phenomenon of *overfitting*, common failure mode of empirical risk minimization.

Generalization: we want rules to perform well on new data points that often are different from those in the training set.
Overfitting: good performance on training data but poor generalization.

How do we fix this problem? Two common solutions: either restrict the set of possible h or appproximate the distribution of (X, Y) differently. We will start with the second strategy, often referred to as *modeling*, and the technique of *maximum likelihood estimation*.

2 Maximum Likelihood Estimation

Modeling: We want to approximate the distribution of the data. For simplicity for the moment, we will consider data with no features, but analogous arguments will also apply once we reintroduce the features X.

Let's assume $y^{(n)}$ are i.i.d. draws from a distribution indexed by a parameter $\theta \in \Theta$. $p(y|\theta)$ is the density or pmf, but when looked as a function of θ, it is often called the likelihood.

If the parameter is finite-dimensional we call it a parametric model

Maximum Likelihood Estimation (MLE): we approximate the distribution as $p(y|\hat{\theta})$ where $\hat{\theta}$ is chosen to maximize the likelihood of the training data.

Example (Bernoulli): $y^{(n)} \in \{0, 1\}$ is iid sampled from some Bernoulli(θ) distribution with $\theta \in [0, 1]$. What is the MLE estimate $\hat{\theta}$?

Because we assumed these points are independent (conditionally independent when working with some features X):

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} p(y^{(n)}|\theta) = \prod_{n=1}^{N} \theta^{y^{(n)}} (1-\theta)^{1-y^{(n)}}$$

Finding the maximum of the likelihood is equivalent to finding the maximum of the log-likelihood as the logarithm is a monotonically increasing function, moreover working in log-space often makes calculations much easier transforming products in sums, therefore we often prefer working with log-likelihoods. This is also true when solving problems computationally where likelihood values become so small that they often vanish with the finite precision of floating point representations. Note that technically taking the log requires us to make sure θ is not 0 or 1, cases which one would have to do separately. In log space:

$$\log p(\mathcal{D}|\theta) = \sum_{n=1}^{N} [y^{(n)} \log \theta + (1 - y^{(n)}) \log(1-\theta)]$$

To find the maximum we show that the second derivative is always non-positive and therefore the maximum has to be to either one of the two extremes (0 or 1) or at a value where the first derivative is 0:

$$\frac{d \log p(\mathcal{D}|\theta)}{d\theta} = \theta^{-1} \sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-1} \sum_{n=1}^{N} (1 - y^{(n)})$$

$$\frac{d^2 \log p(\mathcal{D}|\theta)}{d\theta^2} = -\theta^{-2} \sum_{n=1}^{N} y^{(n)} - (1-\theta)^{-2} \sum_{n=1}^{N} (1 - y^{(n)})$$

All the coefficients and the terms of the sums are non-negative so the second derivative is always non-positive, setting the first derivative equals to 0, we obtain:

$$\hat{\theta} = \arg \max_{\theta \in [0,1]} \log p(\mathcal{D}|\theta) = N^{-1} \sum_{n=1}^{N} y^{(n)}$$

Exercise: Show that for a normally distributed random variable $y^{(n)} \in \mathbb{R}, y^{(n)} \sim \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$, the MLE is $\hat{\mu} = N^{-1} \sum_{n=1}^{N} y^{(n)} =: \bar{y}$ and $\hat{\sigma}^2 = N^{-1} \sum_{n=1}^{N} (y^{(n)} - \bar{y})^2$.

Advantages of MLE:

1. If the likelihood is differentiable w.r.t. the parameters, it is often easy and fast to use modern gradient-based optimizers and complex models. Although in theory these may get stuck in local minima, this scheme is at the heart of many successful uses of ML in practice.

2. MLE is invariant to reparameterization using a bijective function $\eta = f(\theta)$ (as an example using σ^2 instead of σ as parameter).

   Proof: Assuming the MLE is unique:

$$\begin{aligned}
&\forall \theta \neq \hat{\theta}, p(\mathcal{D}|\hat{\theta}) > p(\mathcal{D}|\theta) \\
&\forall \theta \neq \hat{\theta}, p(\mathcal{D}|f^{-1}(f(\hat{\theta}))) > p(\mathcal{D}|f^{-1}(f(\theta))) \\
&\forall \eta \neq f(\hat{\theta}), \tilde{p}(\mathcal{D}|f(\hat{\theta})) > \tilde{p}(\mathcal{D}|\eta) \implies \hat{\eta} = f(\hat{\theta})
\end{aligned}$$

   where the last line follows from a change in parameter which, unlike changes in random variables, does not require a Jacobian term.

Issues with MLE:

1. **Lack of uncertainty:** the output of the MLE does not provide any information on the uncertainty involved in the determination of the estimator itself.

   Example: Modeling a set of datapoints with a normal $N(\mu, 100^2)$ where the mean μ is a parameter. If we have one datapoint $y^{(1)} = 0.5$, the likelihood curve will be very flat as there is little evidence for precise values within one standard deviation of the data. On the other hand, if we have one million data points with a mean of 0.5, the likelihood curve will be steep around $\mu = 0.5$ as values different from it have significantly lower likelihood under the observed data. However, in both cases, the resulting model coming out of the MLE is that the data has a distribution of $\mathcal{N}(y|\hat{\mu} = 0.5, 100^2)$.

2. **Poor generalization with little data:** when the number of data points is small (compared to the number of parameters) the MLE often does not generalize very well.

   Example: $y^{(n)} \in \{0, 1\}$ is iid sampled from some Bernoulli(θ) distribution with $\theta \in [0, 1]$. We have seen that: $\hat{\theta} = \arg\max_{\theta \in [0,1]} \log p(\mathcal{D}|\theta) = N^{-1} \sum_{n=1}^{N} y^{(n)}$. If we have a small amount of data and have not seen any 1's yet, our MLE will put zero likelihood into any 1's coming in the future. A prediction that seems unlikely to generalize well.

3. **Arbitrary likelihood edge cases:** for many models, especially in continuous space, one may be able to obtain arbitrarily large likelihood with extreme parameter settings that overfit specific training data points.

   Example: We have a mixture of two Gaussians with fixed proportions $\pi_k \in (0,1), y^{(n)} \in \mathbb{R}, p(y|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2) = \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$. We will construct a parameter setting with arbitrarily large likelihood, which, by definition, will be among the MLE. Set $\mu_1^* = y^{(1)}$, then

$$p(y^{(1)}|\mu_1^*,\mu_2,\sigma_1^2,\sigma_2^2) \geq \pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} \quad \text{and}$$

$$p(\mathcal{D}|\mu_1^*,\mu_2,\sigma_1^2,\sigma_2^2) \geq \left(\pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}}\right) \prod_{n=2}^{N} \pi_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(y^{(n)}-\mu_2)^2}{2\sigma_2^2}\right)$$

   where the inequality follows from using the iid assumption, writing likelihoods with the sum of Gaussian densities and only keeping one of the two Gaussian for each term (the densities are non-negative). However, now, fixing some values for μ_2 and σ_2, we can make the likelihood arbitrarily large by making σ_1 arbitrarily small. This setting will therefore be a MLE but, once again, it seems unlikely to generalize well.