Recitation 2: Nov 2

Lecturer: Regev Schweiger *Scribe: Yishay Mansour*

2.1 Cross Validation

Cross validation is a method to test the performance of your classifier when there is a limited amount of data available.

We have as an input a set of examples S. We have an algorithm that given a sample T generates a hypothesis h_T (a suggested classifier).

In the cross validation method, we will partition the sample *randomly* to k equal-size parts. Let $S_1, \ldots, S_k$ be the partition. We will run k iterations of our learning algorithm, where in iteration i we have as input $S - S_i$, and compute a hypothesis h_i. We test the hypothesis h_i on S_i and compute its observed error $error_i$. Our prediction of the error of our hypothesis would be the average of the observed errors, i.e., $\frac{1}{k}\sum_{i=1}^{k} error_i$.

2.2 Maximum Likelihood

Consider a Poisson distribution. A Poisson distribution is defined by a parameter $\lambda > 0$ and the probability is defined over the integers and denoted by $Pois(\lambda)$. The motivation is that it models an arrival rates of individuals with an average arrival rate of λ. The probability of having k individual arrive when $X \sim Pois(\lambda)$ is,

$$\Pr[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Assume we have a sample of n points $S = \{z_i, \ldots, z_n\}$ where each z_i is drawn independently from a distribution $Pois(\lambda)$. The likelihood function would be,

$$L_S(\lambda) = \Pr[S|\lambda] = \prod_{i=1}^{n} \Pr[z_i|\lambda] = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda_i^z}{z_i!}.$$

Many times, it is more convenient to work with the *log-likelihood*, simply taking the logarithm of the likelihood, and the product becomes a sum. Note that maximizing the likelihood is equivalent to maximizing the log-likelihood. In our case, the log-likelihood is:

$$\ell_S(\lambda) = \log L_S(\lambda) = \sum_{i=1}^{n} (-\lambda + z_i \log \lambda - \log(z_i!))$$

We would like to find the λ that maximizes the likelihood, denoted by λ_{ML}. Since the terms $\log(z_i!)$ do not depend on λ we can ignore them in the maximization. We have,

$$\lambda_{ML} = \arg\max_{\lambda} \left(-n\lambda + (\sum_{i=1}^{n} z_i) \log \lambda \right)$$

Taking the derivative and equating with zero we have,

$$0 = -n + \left(\sum_{i=1}^{n} z_i \right) \frac{1}{\lambda_{ML}}$$

and the solution is,

$$\lambda_{ML} = \frac{\sum_{i=1}^{n} z_i}{n}.$$

We need to verify that this is indeed a maximum. The second derivative is

$$\left(\sum_{i=1}^{n} z_i \right) \frac{-1}{\lambda^2} < 0$$

and therefore we found a maximum.

2.3 Naïve Bayes

In the problem of classification, we are given a set of samples $(\mathbf{x}_i, y_i)$, drawn from a joint distribution. We want to learn a classifier $f(\mathbf{x}) \approx y$. Often, the set of all possible distributions is too large, so we are forced to make simplifying assumptions about it. We hope that the assumptions are not too far from the truth, and that they will allow us to effectively build a good classifier.

In the case of Naïve Bayes, the assumption is that each of the features is independent of the other features, conditioned on the class:

$$\Pr[\mathbf{x}, y] = \Pr[y] \cdot \Pr[\mathbf{x}|y] = \Pr[y] \cdot \prod_{i=1}^{d} \Pr[x^i|y]$$

If we know (or learn) the distribution, then a good classifier might be to select the most likely class given the data. Namely,

$$h(\mathbf{x}) = \arg\max_{y \in C} \Pr[y|\mathbf{x}] = \arg\max_{y \in C} \frac{\Pr[y, \mathbf{x}]}{\Pr[\mathbf{x}]}$$

Since $\Pr[\mathbf{x}]$ does not depend on the class $y \in C$, we can ignore it and have

$$\begin{aligned} h(\mathbf{x}) &= \arg\max_{y \in C} \Pr[y, \mathbf{x}] \\ &= \arg\max_{y \in C} \Pr[y] \Pr[\mathbf{x}|y] \\ &= \arg\max_{y \in C} \log \Pr[y] + \log \Pr[\mathbf{x}|y] \end{aligned}$$

the last identity follows since the logarithm is a monotone increasing function, hence taking log does not change the maximization problem. We want to model $\Pr[\mathbf{x}|y]$. The Naïve Bayes assumption is that given the class y, the d attributes in $\mathbf{x}$ are independent. Namely,

$$\Pr[\mathbf{x}|y] = \prod_{j=1}^{d} \Pr[x^j|y]$$

This implies that in the maximization we have

$$h(\mathbf{x}) = \arg\max_{y \in C} \log \Pr[y] + \sum_{j=1}^{d} \log \Pr[x^j|y]$$

Assume we ask $1,000$ people about their radio listening habits. Each person specifies whether he or she listens to network A, to network B and to network C. (The feedback is Boolean, so we have three Boolean attributes for each person.) In addition each person is asked if their income above or below the average. (We denote by 1 above the average and by 0 below the average.)

This implies that our sample is $S = \{\mathbf{z}_i\}_{i=1}^{1000}$ where $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ and $\mathbf{x}_i \in \{0,1\}^3$, telling which network a person listens to, and $y_i \in \{0,1\}$ is the indicator whether the salary of the person is above (1) or below (0) the average.

Consider the following classification goal: *Given the listening preferences of a person, decide if their salary is above or below average.*

Let's consider it more abstractly. Assume we have a set of possible outcomes C. (In our example $C = \{0,1\}$.) We have d Boolean attributes for each example (in the example $d = 3$). As our prediction, we like to select the class $y \in C$ which is most likely given the observation $\mathbf{x}$.

The main point is that we can estimate each of the parameters easily from the data. One way of doing the estimate is considering them as a product of Bernoulli variables. The maximum likelihood for each variable in this case would be the empirical frequency (as shown in the lecture).

In our example, the model includes:

$$\begin{aligned}
\theta_1 &= \Pr[y=1] \\
(\theta_0 &= \Pr[y=0] = 1-\theta_1) \\
\theta^i_{1|0} &= \Pr[x^i=1|y=0] \\
(\theta^i_{0|0} &= \Pr[x^i=0|y=0] = 1-\theta^i_{1|0}) \\
\theta^i_{1|1} &= \Pr[x^i=1|y=1] \\
(\theta^i_{0|1} &= \Pr[x^i=0|y=1] = 1-\theta^i_{1|1})
\end{aligned}$$

Which gives a total of $1+2d$ parameters (compared to $2^{d+1}-1$ parameters of the full distribution). Let $\#(I)$ be the number of records that have property I. Using the Maximum Likelihood (ML) for Bernoulli variables, we get:

$$\begin{aligned}
\widehat{\theta}_0 &= \frac{\#(y=0)}{n}, \\
\widehat{\theta}^i_{1|0} &= \frac{\#(x^i=1, y=0)}{\#(y=0)}, \\
\widehat{\theta}^i_{1|1} &= \frac{\#(x^i=1, y=1)}{\#(y=1)},
\end{aligned}$$