Recitation 2: October 20

Lecturer: Mariano Schain *Scribe: ym*

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.

In the cross validation we will partition the sample *randomly* to k equal size parts. Let $S_1, \ldots, S_k$ be the partition. We will run k iteration 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 define over integers and denoted by $Pois(\lambda)$. The motivation is that it models an arrival rates of individual 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,

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

It is many times more convenient to work with the log-Likelihood, simply taking the logarithm of the likelihood, and the product becomes a sum. Note that if maximizing the likelihood, it will be equivalent to maximizing the log-likelihood.

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

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,

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

Taking the derivative and equating with zero we have,

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

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

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

and therefore we found a maximum.

2.3 Naïve Bayes

Assume we ask $1,000$ people about there radio listening habits. Each specifies whether he 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 his income above or below the average. (We denote by A above the average and by B below the average.)

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

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

Lets consider it more abstractly. Assume we have a set of possible outcomes C. (In our example $C = \{A, B\}$.) We have d Boolean attributes for each example (in the example $d = 3$). As our prediction, we like to select the class $c \in C$ which is most likely given the observation x. Namely,

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

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

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

the last identity follows since the logarithm is a monotone increasing function, hence taking log does not change the maximization problem.

Now we get to the point that we want to model $\Pr[x|C = c]$. The Naive Bayes assumption is that given the class $C = c$ the d attributes in x are independent. Namely,

$$\Pr[x | C = c] = \prod_{j=1}^{d} \Pr[x^j | C = c]$$

This implies that in the maximization we have

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

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 Bernoulli variable. The maximum likelihood in this case would be the empirical frequency (as shown in the lecture).

Back to our example. The model there includes

$$(\theta_A, \theta_B, \{\theta_{j,A}, \theta_{j,B}\}_{j=1}^{3}),$$

where $\theta_A = \Pr[C = A]$, $\theta_B = \Pr[C = B]$, $\theta_{j,A} = \Pr[x^j = 1 | C = A]$ and $\theta_{j,B} = \Pr[x^j = 1 | C = B]$.

Let $\#(I)$ be the number of records that have property I.

Using the Maximum Likelihood (ML) we set:

$$\begin{aligned} \widehat{\theta}_A &= \frac{\#(c_i = A)}{n}, \\ \widehat{\theta}_B &= \frac{\#(c_i = B)}{n} = 1 - \theta_A, \\ \theta_{j,A} &= \frac{\#(x_i^j = 1, c_i = A)}{\#(c_i = A)} \\ \theta_{j,B} &= \frac{\#(x_i^j = 1, c_i = B)}{\#(c_i = B)} \end{aligned}$$