CPSC 540 Notes on Naive Bayes

Mark Schmidt

Fall 2014

The maximum likelihood estimator (MLE) is the hypothesis h that maximizes the likelihood, $p(\mathcal{D}|h)$, of a dataset $\mathcal{D}$ over a set of possible hypotheses $\mathcal{H}$,

$$h_{\text{MLE}} = \arg\max_{h\in\mathcal{H}} p(\mathcal{D}|h).$$

In supervised learning, we define $\mathcal{D}$ as a set of ordered pairs $\{(x_i, y_i)\}_{i=1}^N$, and we'll use $\mathcal{D}_i$ to denote ordered pair number i, (x_i, y_i). If we assume these ordered pairs are independent and identically distributed (IID), we have

$$h_{\text{MLE}} = \arg\max_{h\in\mathcal{H}} \prod_{i=1}^N p(\mathcal{D}_i|h).$$

Generative classifiers model the probability $p(y_i, x_i|h)$,

$$p(\mathcal{D}_i|h) = p(y_i, x_i|h) = p(x_i|y_i, h)p(y_i|h).$$

In naive Bayes, we assume that the variables $x^1, x^2, \ldots, x^D$ are mutually conditionally independent given y, which gives us

$$\begin{aligned}
p(x_i|y_i, h) &= p(x_i^1|x_i^{2:D}, y_i, h)p(x_i^{2:D}|y_i, h)\\
&= p(x_i^1|y_i, h)p(x_i^{2:D}|y_i, h)\\
&= p(x_i^1|y_i, h)p(x_i^2|x_i^{3:D}, y_i, h)p(x_i^{3:D}|y_i, h)\\
&= p(x_i^1|y_i, h)p(x_i^2|y_i, h)p(x_i^{3:D}|y_i, h)\\
&= \prod_{j=1}^D p(x_i^j|y_i, h)
\end{aligned}$$

We need to choose how we will define $p(y_i|h)$ and $p(x_i^j|y_i, h)$. If y is binary $\{0, 1\}$, then it makes sense to use Bernoulli distributions (If we toss a coin that lands 'heads' with probability θ, we say that the distribution of {heads,tails} follows a Bernoulli distribution with parameter θ).

We'll use θ as the parameter of the Bernoulli distribution for y_i, so that y_i is distributed according to a Bernoulli random variable with parameter θ (so we'll have $\theta \in h$, and h will also include the parameters of the other distributions we'll use in the model), which we write as

$$y_i \sim \text{Ber}(\theta),$$

From the definition of a Bernoulli random variable, we have under this assumption that

$$p(y_i|h) = p(y_i|\theta) = \theta^{I(y_i=1)}(1-\theta)^{I(y_i=0)}.$$

If the x_i^j are also binary, it still makes sense to use a Bernoulli distribution but we will have a different Bernoulli distribution depending on the value of the corresponding y_i. So for each variable j we will have two parameters θ_1^j and θ_0^j, and the value of y_i decides which one we use,

$$x_i^j | y_i \sim \mathrm{Ber}(\theta_{y_i}^j),$$

$$p(x_i^j | y_i, h) = p(x_i^j | y_i, \theta_{y_i}^j) = (\theta_{y_i}^j)^{I(x_i^j=1)}(1-\theta_{y_i}^j)^{I(x_i^j=0)}.$$

To compute the MLE, for numerical reasons we typically work in the log-domain (taking the logarithm doesn't change the argmax), and plugging in everything above we get

$$\begin{aligned}
h_{\mathrm{MLE}} &= \arg\max_{h\in\mathcal{H}} p(\mathcal{D}|h) && \text{(definition of MLE)}\\
&= \arg\max_{h\in\mathcal{H}} \prod_{i=1}^N p(\mathcal{D}_i|h) && \text{(IID assumption)}\\
&= \arg\max_{h\in\mathcal{H}} \log \prod_{i=1}^N p(\mathcal{D}_i|h) && \text{(log does not change optimal value)}\\
&= \arg\max_{h\in\mathcal{H}} \sum_{i=1}^N \log p(\mathcal{D}_i|h) && \text{(log turns multiplication in addition)}\\
&= \arg\max_{h\in\mathcal{H}} \sum_{i=1}^N \log p(y_i, x_i|h) && \text{(definition of } \mathcal{D}_i\text{)}\\
&= \arg\max_{h\in\mathcal{H}} \sum_{i=1}^N \log(p(y_i|h)p(x_i|y_i,h)) && \text{(product rule)}\\
&= \arg\max_{h\in\mathcal{H}} \sum_{i=1}^N [\log p(y_i|h) + \log p(x_i|y_i,h)] && \text{(log turns multiplication into addition)}\\
&= \arg\max_{h\in\mathcal{H}} \sum_{i=1}^N \left[\log p(y_i|h) + \sum_{j=1}^D \log p(x_i^j|y_i,h)\right] && \text{(naive Bayes assumption)}\\
&= \arg\max_{\theta\in[0,1],\theta_i^j\in[0,1],\forall_{i\in\{0,1\},j\in\{1,2,\ldots,D\}}} \sum_{i=1}^N \left[\log p(y_i|\theta) + \sum_{j=1}^D \log p(x_i^j|y_i,\theta_{y_i}^j)\right] && \text{(Bernoulli parameterization)}
\end{aligned}$$

Each term in this sum only depends on either θ or a single value of θ_1^j or θ_0^j. This means we can solve for these parameters independently (in optimizaiton, this is called a *separable* function). Let's just concentrate on the terms that depend on θ:

$$
\begin{aligned}
\theta_{MLE} &= \arg\max_{\theta \in [0,1]} \sum_{i=1}^{N} \log p(y_i|\theta) \\
&= \arg\max_{\theta \in [0,1]} \sum_{i=1}^{N} \log(\theta^{I(y_i=1)}(1-\theta)^{I(y_i=0)}) \\
&= \arg\max_{\theta \in [0,1]} \sum_{i=1}^{N} [I(y_i=1)\log(\theta) + I(y_i=0)\log(1-\theta))] \\
&= \arg\max_{\theta \in [0,1]} \log(\theta) \sum_{i=1}^{N} I(y_i=1) + \log(1-\theta) \sum_{i=1}^{N} I(y_i=0) \\
&= \arg\max_{\theta \in [0,1]} \log(\theta) N_1 + \log(1-\theta) N_0,
\end{aligned}
$$

where N_1 is the number of times $y_i = 1$ in the training data and N_0 is the number of times $y_i = 0$. We will be able to prove this with tools we develop later, but right now I will claim that there is one stationary point of the log-likelihood in terms of θ in the interval $[0, 1]$ and that this is a maximizer. To find this stationary point, take the derivative and set it to 0,

$$0 = \frac{N_1}{\theta} - \frac{N_0}{1-\theta}.$$

Re-arrange this to get

$$\frac{\theta}{1-\theta} = \frac{N_1}{N_0} = \frac{N_1/N}{N_0/N}.$$

The solution to this (within $[0, 1]$) is

$$\theta = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N},$$

to see this observe that $1 - \theta = 1 - N_1/N = N_0/N$.

This is an overly complicated way to say that if you flip a coin 100 times and it lands heads 40 times, then if you have no prior knowledge your most likely guess for the probability that it will land heads is 40/100.

The general solution when $y_i \in \{1, 2, \ldots, C\}$ and we have parameters $\{\theta_1, \theta_2, \ldots, \theta_C\}$ is

$$\theta_c = \frac{N_c}{N}.$$

For a binary x_i conditioned on these y_i, you get

$$\theta_c^j = \frac{N_{c1}^j}{N_c},$$

where N_{c1}^j is the number of times variable $x_i^j = 1$ and $y_i = c$.