Introduction to Machine Learning **Fall Semester**

Recitation 5

Lecturer: Regev Schweiger *Scribe: Regev Schweiger*

5.1 Naïve Bayes

In the problem of classification, we are given a set of samples $(\mathbf{x}_i, y_i)$, drawn from a joint distribution of the random variables $(\mathbf{X}, Y)$. 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.

As an example, assume we test 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 age is above or below thirty. (We denote by 1 above thirty and by 0 below thirty.)

In this example, 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 age of the person is above (1) or below (0) thirty. Consider the following classification goal: *Given the listening preferences of a person, decide if their age is above or below thirty.*

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

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

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

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

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 $\Pr[\mathbf{X} = \mathbf{x} | Y = y]$.

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

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

This implies that in the maximization we have

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

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:

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

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:

$$\begin{aligned}
\widehat{\theta}_0 &= \frac{\#(Y = 0)}{n}, \\
\widehat{\theta}^j_{1|0} &= \frac{\#(X^j = 1, Y = 0)}{\#(Y = 0)}, \\
\widehat{\theta}^j_{1|1} &= \frac{\#(X^j = 1, Y = 1)}{\#(Y = 1)}
\end{aligned}$$

5.2 Constrained Optimization

5.2.1 Lagrange Multipliers

Suppose we have the following problem of constrained optimization with equality constraints:

$$\begin{aligned} &\min_{\mathbf{x}\in \mathrm{R}^d} f(\mathbf{x}) \\ &\text{s.t.} \\ &g_i(\mathbf{x}) = 0 \ \text{ for } i = 1, \ldots, N \end{aligned}$$

Define the *Lagrangian* of this problem as follows:

$$\mathcal{L}(\mathbf{x}, \lambda_1, \ldots, \lambda_N) = f(\mathbf{x}) + \lambda_1 g_1(\mathbf{x}) + \ldots + \lambda_N g_N(\mathbf{x})$$

Under some regularity requirements, a necessary condition for $\mathbf{x}^*$ being a critical point for f under the constraints, it that there exist values $\lambda_1^*, \ldots, \lambda_N^*$, such that $(\mathbf{x}^*, \lambda_1^*, \ldots, \lambda_N^*)$ is a critical point of $\mathcal{L}$ (unconstrained); and specifically, that the gradient of the Lagrangian at the point is 0:

$$\nabla \mathcal{L}(\mathbf{x}^*, \lambda_1^*, \ldots, \lambda_N^*) = \mathbf{0}$$

We will not prove this here.

5.2.2 Karush-Kuhn-Tucker Conditions

The KKT conditions extend the ideas of Lagrange multipliers to handle inequality constraints in addition to equality constraints. While there is a general formulation including both equalities and inequalities, we will limit ourselves to provide a first-order optimality condition for the problem:

$$\begin{aligned} &\min_{\mathbf{x}\in \mathrm{R}^d} f(\mathbf{x}) \\ &\text{s.t.} \\ &g_i(\mathbf{x}) \leq 0 \ \text{ for } i = 1, \ldots, N \end{aligned}$$

The Lagrangian is again defined as

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) = f(\mathbf{x}) + \sum_{i=1}^{N} \alpha_i g_i(\mathbf{x})$$

where $\boldsymbol{\alpha}$ are called the *Lagrangian multipliers*. The theory (on which we do not elaborate here) tells us that the dual program:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}) \\ &\quad \text{s.t.} \;\; \alpha_i \geq 0 \qquad \forall i = 1, \ldots, N \end{aligned}$$

achieves the same optimal point for all the cases which we will consider. Meaning, that if $\boldsymbol{\alpha}^*$ is a solution for the dual problem, then $\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\alpha}^*)$ is a solution for the original problem (under some conditions, that we will assume to hold). As an example, let us solve the problem

$$\begin{aligned} &\min_{(x_1, x_2) \in \mathrm{R}^2} x_1^2 + x_2^2 \\ &\quad \text{s.t.} \\ &x_1 + x_2 \geq 2 \end{aligned}$$

The Lagrangian is:

$$\mathcal{L}(x_1, x_2, \boldsymbol{\alpha}) = x_1^2 + x_2^2 + \alpha(-x_1 - x_2 + 2)$$

For a fixed $\boldsymbol{\alpha}$, we want to calculate $\min_{(x_1, x_2)} \mathcal{L}(x_1, x_2, \boldsymbol{\alpha})$. To do that, we equate the derivative to zero:

$$\frac{\partial \mathcal{L}}{\partial x_i} = 2x_i - \alpha = 0 \Rightarrow x_i^* = \alpha/2$$

the second derivative is positive, so this is indeed a minimum. The value of the minimum is:

$$\mathcal{L}(x_1^*, x_2^*, \alpha) = 2 \cdot \frac{\alpha^2}{4} + \alpha \left(-2 \cdot \frac{\alpha}{2} + 2 \right) = 2\alpha - \frac{\alpha^2}{2}$$

Finding the maximum over α, subject to $\alpha \geq 0$, gives $\alpha^* = 2$. Therefore, the optimal points are $x_1^* = x_2^* = 1$.