Chris Piech
CS109

Section #8
May 29-31, 2019

Section #8: Machine Learning Soln

1. **Vision Test**:

We are going to solve this problem by finding the MLE estimate of θ. To find the MLE estimate, we are going to find the argmax of the log likelihood function. To calculate argmax we are going to use gradient ascent, which requires that we know the partial derivative of the log likelihood function with respect to theta.

First we write the log likelihood:

$$L(\theta) = \prod_{i=1}^{20} p^{y^{(i)}}(1-p)^{[1-y^{(i)}]}$$
$$LL(\theta) = \sum_{i=1}^{20}(y^{(i)}\log(p) + (1-y^{(i)})\log(1-p))$$

Then we find the derivative of log likelihood with respect to θ. We first do this for one data point:

$$\frac{\partial LL}{\partial \theta} = \frac{\partial LL}{\partial p} \cdot \frac{\partial p}{\partial \theta}$$

We can calculate both the smaller partial derivatives independently:

$$\frac{\partial LL}{\partial p} = \frac{y^{(i)}}{p} - \frac{1-y^{(i)}}{1-p}$$
$$\frac{\partial p}{\partial \theta} = p[1-p]$$

Putting it all together for one letter:

$$\begin{aligned}
\frac{\partial LL}{\partial \theta} &= \frac{\partial LL}{\partial p} \cdot \frac{\partial p}{\partial \theta} \\
&= [\frac{y^{(i)}}{p} - \frac{1-y^{(i)}}{1-p}]p[1-p] \\
&= y^{(i)}(1-p) - p(1-y^{(i)}) \\
&= y^{(i)} - p \\
&= y^{(i)} - \sigma(\theta - f)
\end{aligned}$$

For all twenty examples:

$$\frac{\partial LL}{\partial \theta} = \sum_{i=1}^{20} y^{(i)} - \sigma(\theta + f^{(i)})$$

2. **Multiclass Bayes**: We can make the Naive Bayes assumption of independence and simplify argmax of $P(Y|\mathbf{X})$ to get an expression for $\hat{Y}$, the predicted output value, and evaluate it using the provided **count** function.

$$
\begin{aligned}
\hat{Y} &= \arg\max_y \frac{P(X_1 = 1, X_2 = 1, X_3 = 0|Y = y)P(Y = y)}{P(X_1 = 1, X_2 = 1, X_3 = 0)} \\
&= \arg\max_y P(X_1 = 1, X_2 = 1, X_3 = 0|Y = y)P(Y = y) \\
&= \arg\max_y P(X_1 = 1|Y = y)P(X_2 = 1|Y = y)P(X_3 = 0|Y = y)P(Y = y), \text{ where:}
\end{aligned}
$$

$$
\begin{aligned}
P(X_1 = 1|Y = y) &= [\mathbf{count}(X_1 = 1, Y = y) + 1]/\mathbf{count}(Y = y) + 2 \\
P(X_2 = 1|Y = y) &= [\mathbf{count}(X_2 = 1, Y = y) + 1]/\mathbf{count}(Y = y) + 2 \\
P(X_3 = 1|Y = y) &= [\mathbf{count}(X_3 = 1, Y = y) + 1]/\mathbf{count}(Y = y) + 2 \\
P(X_1 = 0|Y = y) &= [\mathbf{count}(X_1 = 0, Y = y) + 1]/\mathbf{count}(Y = y) + 2 \\
P(X_2 = 0|Y = y) &= [\mathbf{count}(X_2 = 0, Y = y) + 1]/\mathbf{count}(Y = y) + 2 \\
P(X_3 = 0|Y = y) &= [\mathbf{count}(X_3 = 0, Y = y) + 1]/\mathbf{count}(Y = y) + 2
\end{aligned}
$$

and you don't need to use MAP to estimate Y:
$P(Y = y) = \mathbf{count}(Y = y)/10,000$

3. **The Most Important Features**

a. $LL(\theta) = y \cdot \log \sigma\left(\theta^T \cdot \mathbf{x}\right) + \left(1 - y\right)\log\left[1 - \sigma\left(\theta^T \cdot \mathbf{x}\right)\right]$

b. We can calculate the saliency for a single feature as follows.

$$
\begin{aligned}
LL(\theta) &= y \log z + \left(1 - y\right) \log \left(1 - z\right) && \text{where } z = \sigma\left(\theta^T \cdot \mathbf{x}\right) \\
\frac{\partial LL}{\partial x_i} &= \frac{\partial LL}{\partial z} \cdot \frac{\partial z}{\partial x_i} && \text{chain rule} \\
&= \left(\frac{y}{z} - \frac{1 - y}{1 - z}\right) \cdot \left(z(1 - z)\theta_i\right) && \text{partial derivatives} \\
\text{saliency} &= \left| \left(\frac{y}{z} - \frac{1 - y}{1 - z}\right)z(1 - z)\theta_i \right|
\end{aligned}
$$

c. We can take the ratio as follows using our expression above.

$$
\text{saliency for feature } i, S_i = \left| \left(\frac{y}{z} - \frac{1 - y}{1 - z}\right)z(1 - z)\theta_i \right|, \text{ and same for } S_j
$$

$$
\frac{S_i}{S_j} = \frac{\left| \left(\frac{y}{z} - \frac{1-y}{1-z}\right)z(1 - z)\theta_i \right|}{\left| \left(\frac{y}{z} - \frac{1-y}{1-z}\right)z(1 - z)\theta_i \right|} = \frac{S_i}{S_j} = \frac{\left|\theta_i\right|}{\left|\theta_j\right|} \text{ by elimination}
$$