Logistic Regression

Recap

Last time we considered how to make a hypothesis that approximates a real value function. All we need to do is build a cost function and then minimize it. This time we are going to use the same idea for binary function.

Logistic Regression

Suppose you have data points

$$(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)}), \dots (\mathbf{x}^{(N)}, y^{(N)})$$

where $\mathbf{x}^{(i)}$ are feature vectors and $y^{(i)} \in \{+1, -1\}$ are the classes.

Now the question is the input feature $\mathbf{x}$ what is the probability that this datapoint would be of class $+1$.

So, the unknown target function in this case is a probability. Specifically,

$$f(\mathbf{x}) = \mathrm{P}[y = +1|\mathbf{x}]$$

The probability on the right reads: the probability that the given data point is of class $+1$ given that the data point has input feature vector $\mathbf{x}$. For example, the probability that a person would have more than 50k income given that he is age 35, working in computer industry, married and have 2 kids.

Note that this also imply

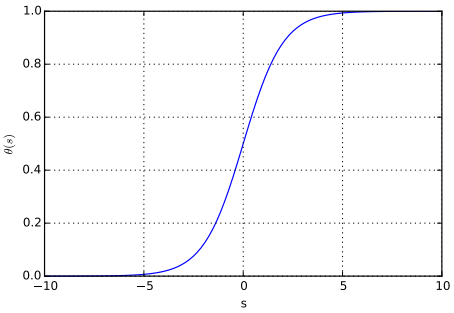
$$\mathrm{P}[y = -1|\mathbf{x}] = 1 - f(\mathbf{x})$$

The probability is a real number between 0 and 1. We can't use linear regression for this problem since the probability ranges from +1 to -1. But the output of the linear regression is from $-\infty$ to ∞. First attempt to fix this could be trying to take the sign of the linear combination and shift it up then scale it so it ranges from -1 to 1. The problem with this approach is that there would be no input vector taking the value in between 0 and 1; ex: 0.8.

To solve this issue we need to introduce a soft threshold function. There many soft threshold function but the most popular one is called logistic function

$$\theta(s) = \frac{\mathrm{e}^s}{1 + \mathrm{e}^s} = \frac{1}{\mathrm{e}^{-s} + 1}$$

All this function does is that it convert s from $-\infty$ to ∞ to 0 and 1 continuously. So we have a chance of making it probability. The second one is more friendly to computer round off.



Before we go on there is once property of θ that makes this particularly useful.

$$\theta(-s) = 1 - \theta(s)$$

You can verify this easily.

Let us use this to make hypothesis for the probability of a data point x being of class $+1$:

$$h_{\mathbf{w}}(\mathbf{x}) = \theta(\mathbf{w}^T\mathbf{x}) = \frac{\mathrm{e}^{\mathbf{w}^T\mathbf{x}}}{1 + \mathrm{e}^{\mathbf{w}^T\mathbf{x}}}$$

Here I also use $x_0 = 1$ convention to get the threshold term. To interpret this quantity as conditional probability we can do

$$\begin{aligned}\mathrm{P}(y|\mathbf{x}) &= \begin{cases}\theta(\mathbf{w}^T\mathbf{x}) & y = +1\\ 1 - \theta(\mathbf{w}^T\mathbf{x}) & y = -1\end{cases}\\ &= \begin{cases}\theta(\mathbf{w}^T\mathbf{x}) & y = +1\\ \theta(-\mathbf{w}^T\mathbf{x}) & y = -1\end{cases}\\ &= \theta(y\mathbf{w}^T\mathbf{x})\end{aligned}$$

The last line is a very neat thing to have

$$\mathrm{P}(y|\mathbf{x}) = \theta(y\mathbf{w}^T\mathbf{x}) \tag{1}$$

So, now the problem is to find $\mathbf{w}$ so that our hypothesis best represents the conditional probability of the data. To do this, we need to introduce the concept of likelihood. The likelihood is just the probability that you will be given the data set you were given. Specifically,

$$\mathcal{L} = \prod_{i=1}^{N} \mathrm{P}(y^{(i)}|\mathbf{x}^{(i)})$$

If we replace the probability with the hypothesis we have

$$\mathcal{L} = \prod_{i=1}^{N} \theta(y\mathbf{w}^T\mathbf{x})$$

Thus, what we need to do now is the find $\mathbf{w}$ that maximize the likelihood. Let me point out one important caveat. What we are trying to do here is find $\mathbf{w}$ such that the data points we drawn has the highest likelihood. This is NOT the same thing as what is finding $\mathbf{w}$ that the data points we have is most likely drawn from. The second quantity is unknown and can't really be quantified eventhough it is really what we want. But we take the first quantity as being "good enough" for what we are trying to do.

Maximizing the product is quite cumbersome. We can change the product to sum by taking the natural log. Moreover, since most package are written for minimization rather than maixmization; all we need to turn the maximization to minimization problem is just multiplying it by a minus sign. Thus maximizing the likelihood is equivalent to minimizing the following quantity:

$$\begin{aligned}
\mathrm{cost}(\mathbf{w}) &= -\frac{1}{N}\log\mathcal{L} && (2)\\
&= -\frac{1}{N}\sum_{i=1}^{N}\ln\left[\theta(y\mathbf{w}^T\mathbf{x})\right] && (3)\\
&= \frac{1}{N}\sum_{i=1}^{N}\ln\left[\frac{1}{\theta(y\mathbf{w}^T\mathbf{x})}\right] && (4)\\
&= \frac{1}{N}\sum_{i=1}^{N}\ln\left[1+e^{-y\mathbf{w}^T\mathbf{x}}\right] && (5)
\end{aligned}$$

Unlike linear regression, this cost has no closed form solution for $\nabla\mathrm{cost} = 0$. So we need to resort to numerical methods like gradient descent we covered last time.

Once you get the $\mathbf{w}$ that minimize the log likelihood from gradient descent method or others. We will have a hypothesis that approximate the probability of it being class +1.

Hard Threshold

What we did in the last section was all nice that we get an approximation for the target probability but this doesn't tell us whether the new input vector is of the class +1 or -1. This is easy all you need to do is making a cut off.

$$\mathrm{class} = \begin{cases} +1 & h(\mathbf{x}) > c \\ -1 & h(\mathbf{x}) \leq c \end{cases} \tag{6}$$

But, how do we pick the cut off? The answer depends on practical application. It depends on how much you are willing to tolerate false positive vs false negative.

ROC Curve

When you try to make a hard threshold people usually plot false positive rate vs false negative rate for various value of cutoff. This will let you see the performance of your classifier. The fancy name for this curve is called *Receiver Operating Characteristic Curve* (ROC Curve). Three is a whole family of ROC curves; each with different

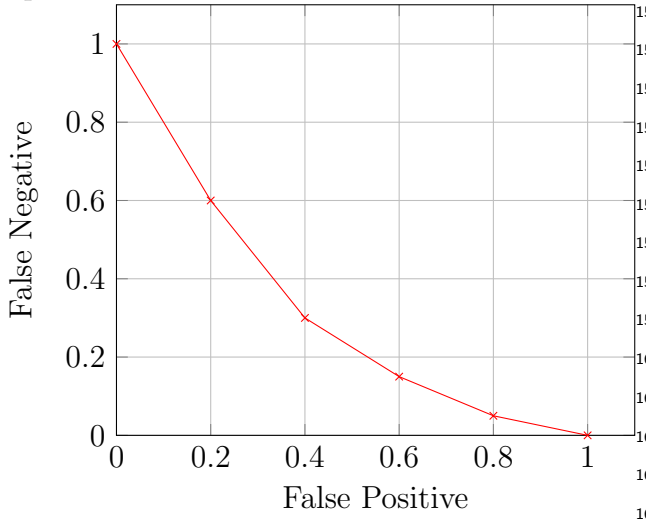
meaning on the axis and different names too, plus each people call it differently too.. Don't bother memorizing the name. Use common sense. The important thing is that when you see one and know what the axis means you know where a good classifier should be and where a bad classifier should be.

For example, if we care about

1. Out of 1000 true positive class how many sample get classified as negative (false negative). We will call this ratio false negative rate.

2. Out for 200 true negative class how many sample get classified as positive (false positive). We will call this ratio false positive rate.

We could plot, for each cutoff we scan over, the false positive rate versus the false negative rate. This will show us how each cut off performs. An example is shown below:



From the figure, let us start out with the lower right corner where it corresponds to cutoff value of zero. At this corner we will never get any false negative since we never classify anything as negative class. The false positive is high since we classify everything as positive. As we increase the cutoff we will classify less thing as positvie and more thing as negative. This will increase false negative and decrease false positive. This will make the curve go toward the upper left. With the cut off really high, you are classifying everything as negative. You will therefore never get false positive since you never classify antying as a positive class to begin with. But you will get 100% false negative.

So if you were to as where we want to be, it is clear that it is lower left corner. This is the place where you have no false positive and no false negative.

Please do not memorize which one increase, which one decrease and where is the perfect place though. You can just use common sense to figure this out from the first principle. There are many variations of ROC curve with different normalization each one will go toward different corner.

For example you can plot the ratio of true positive class classified as positive class versus the ratio of true negative class classified as negative class. This will give you different kind of curve which you can also use to visualize how cutoff affect the performance of the classifier. As an exercise think about which corner a good classifier would be if we plot these two ratio.

The important thing is that when you see one and you konw what each axis represents you can figure out where the good classifer would be.