Conditional Probabilities, Conditional Expectations, and Loss Function

Marc Haddad

> *"The most common reason for not being able to build perfect [predictive] algorithms is that it is impossible"* -Dr. Rafa Irizarry

Observations with the same observed values for predictors may not be the same (e.g. female with height 66 in., and male with height 66 in.; same height, different category).

However, we can assume that the observations have the same probability of being one category or another (e.g. 40% chance of having 66 in. female, 60% chance of having 66 in. male).

Mathematically represented as:

$(X_1 = x_1, X_2 = x_2, ..., X_p = x_p)$

for **observed values** $(x_1, ..., x_p)$ of **covariates** $(X_1, ..., X_p)$.

We denote the conditional probabilities of each class k:

$Pr(Y = k \mid X_1 = x_1,\ X_2 = x_2, ..., X_p = x_p)$
for $k = (1, ..., K)$.

We use bold letters to write out all the predictors like this:

$\mathbf{X} = (X_1, ..., X_p)$ and
$\mathbf{x} = (x_1, ..., x_p)$.

The conditional probability of being in class k is:

$p_k(\mathbf{x}) = Pr\left(Y = k \mid \mathbf{X} = \mathbf{x}\right)$ for $k = (1, ..., K)$.

For any set of predictors $\mathbf{X}$ we will predict the class k with the largest probability among $p_1(x),\ p_2(x), ...,\ p_K(x)$
Which can be written as:

$\hat{Y} = \max_k p_k(\mathbf{x})$

However, we can't compute the above equation because we don't know the p_k of $\mathbf{x}$'s. This exemplifies the main challenge of Machine Learning: Estimating these conditional probabilities.

The better our algorithm estimates $\hat{p_k}(\mathbf{x})$,
the better our predictor $\hat{Y} = \max_k \hat{p_k}(\mathbf{x})$

The quality of our prediction will depend on two things:

1. How close the maximum probability $\max_k p_k(\mathbf{x})$ is to 1
2. How close our estimate of the probabilities $\hat{p_k}(\mathbf{x})$ are to the actual probabilities $p_k(\mathbf{x})$

Because item 1 is determined by the nature of each problem, our best option is to use item 2 to best estimate conditional probabilities.

Though it is our approach here, we must keep in mind that maximizing probability is not always optimal in practice. Our approach depends on context. **Sensitivity** and **Specificity** may differ in importance in different contexts. But having a good estimate of conditional probabilities is more often than not sufficient when building an optimal prediction model due to the fact that we can control both sensitivity and specificity.