Note 5

05 Decision Theory

by Alvin Wan

We now consider a probabilistic classifier. Before we begin, however, let us consider some important concepts from probability. If this is your cup of coffee, feel free to skip to section 3 below.

1 Discrete Probability Review

We know that the joint of two events can be expanded; this is called the **chain rule**.

$$\Pr(A, B) = \Pr(A|B)\Pr(B)$$

We can also sum out over all possibilities of one event, to rid of it. Thus, we obtain the **total probability law**.

$$\Pr(B) = \Pr(A, B) + \Pr(\bar{A}, B) = \Pr(B|A)\Pr(A) + \Pr(B|\bar{A})\Pr(\bar{A})$$

More generally, we have $\Pr(A_1, A_2, \ldots A_n) = \Pr(A_1)\Pr(A_2|A_1)\cdots\Pr(A_n|A_{n-1}, A_{n-2}\ldots, A_1)$. Rearrange the first line above to get our definition of conditional probability.

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

Applying chain rule to the numerator, we have the formulation of **Bayes' rule**:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

Above, $\Pr(A)$ is termed the **prior** and $\Pr(A|B)$ is termed the **posterior**.

2 Continuous Probability Review

In effect, replace all summations with integrals. The expected value of random variable X is defined as

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

where $f_X(x)$ is the **probability density function (pdf)** of X. The **cumulative distribution function (cdf)** $F_X(x)$ is defined to be

$$F_X(x) = \int_{-\infty}^{\infty} f_X(x) dx$$

Variance is defined as the "spread" of our data, just as in discrete probability.

$$\text{var}(X) = E[(X - E[X])^2] = E[X^2 - 2E[X]X + E[X]^2] = E[X^2] - E[X]^2$$

3 Decision Rule

Consider a binary-valued true label Y and binary-valued prediction $\hat{Y}$. Additionally, consider our loss function from before, where $L(\hat{y}, y)$ is 1 if our prediction is not the true label $\hat{y} \neq y$ and 0 otherwise. Now note: If the probability of class 0 is greater than the probability of class 1 given x, $P(Y = -1|x) > P(Y = 1|x)$, we predict class $Y = -1$. This is how our intuitive reasoning works: we pick the **largest posterior probability**. In other words, our decision rule is the following:

$$f(x) = \begin{cases} 1 & \text{if } \Pr(Y = 1|x) > \Pr(Y = -1|x) \\ -1 & \text{otherwise} \end{cases}$$

What if we wish to "weigh" certain misclassifications more? We call $L(1, -1)$ a **false positive**, since we predicted true incorrectly. Call $L(1, -1)$ a **false negative**, since we predicted false incorrectly. In some cases, a false negative may be far worse than a false positive i.e., determining whether a patient has cancer. To formalize this notion, we consider the following **loss matrix**, another representation of our loss function. Here, we heavily penalize a false negative, by setting the value of $L(-1, 1)$ to be arbitrarily high.

	Y=-1	Y=1
$\hat{Y} = -1$	0	100
$\hat{Y} = 1$	1	0

We now consider the expected loss associated with predicting true, $\hat{Y} = 1$.

$$\begin{aligned}
\ell(x|\hat{y} = 1) &= E[L(1, y)] \\
&= \sum_y L(1, y) \Pr(Y = y|x) \\
&= L(1, -1) \Pr(Y = -1|x) + L(1, 1) \Pr(Y = 1|x) \\
&= L(1, -1) \Pr(Y = -1|x)
\end{aligned}$$

We can repeat this to obtain the expected loss associated with predicting false, $\hat{Y} = -1$. We get

$$\ell(x|\hat{y} = -1) = E[L(-1, y)] = L(-1, 1) \Pr(Y = 1|x)$$

Thus, we can construct a new decision rule, where we predict true, if the expected loss for predicting true is less than the expected loss for predicting false. In other words, our prediction function f is the following, which is known as the **Bayes Decision Rule**.

$$f(x) = \begin{cases} 1 & \text{if } \ell(x|\hat{y} = 1) > \ell(x|\hat{y} = -1) \\ -1 & \text{otherwise} \end{cases}$$

Takeaway: If the loss matrix is asymmetric, weight posteriors with losses.

4 Risk Minimization

The above decision rules are built to minimize **risk**, which is total expected loss with respect to all x, y. Taking risk, we have

$$
\begin{aligned}
R(f) &= E[L(f(x), y)] \\
&= \sum_x L(f(x), 1) \Pr(Y = 1, x) + L(f(x), -1) \Pr(Y = -1, x) \\
&= \sum_x L(f(x), 1) \Pr(x|Y = 1) \Pr(Y = 1) + L(f(x), -1) \Pr(x|Y = -1) \Pr(Y = -1)
\end{aligned}
$$

Say that our class-conditional probabilities are continuous densities. Then, we have the following formulation:

$$
\begin{aligned}
R(f) &= E[L(f(x), y)] \\
&= \int_x (L(f(x), 1) f_{X|Y}(x|1) \Pr(Y = 1) + L(f(x), -1) f_{X|Y}(x| - 1) \Pr(Y = -1)) dx
\end{aligned}
$$

Note that the **Bayes optimal decision boundary** is all x such that either class is equally likely, $\{x : \Pr(Y = 1|x) = 0.5\}$. To generalize to more classes, we can simply take the maximum posterior probability, $\max_i \Pr(Y = i|x)$.