Introduction to Decision Theory for Machine Learning

Arvind Sridhar

CS 189/289A, UC Berkeley, Spring 2021

1 Introduction

Decision theory is built around the fundamental concept of **classification**: given an input datapoint X and any number of classes $Y = y_0, y_1, ..., y_n$, under which category should I classify X? In machine learning, this question is often answered by looking at the **loss** encountered when we make an incorrect classification: for example, if we predict that X belongs to class y_3 when in reality it belongs to class y_1, what is the associated penalty that our classifier incurs? If this misclassification is particularly egregious, we might incur a very large penalty; else, the penalty might be small. Therefore, our **training process** involves learning a classifier that **minimizes total penalty incurred on the training dataset**, namely a set of points X for which we know the proper classifications Y. The total penalty on the training dataset can be as simple as the sum of the individual penalties (if any) that the classifier incurs on each point in the training set.

In Bayes decision theory, we view this general classification framework from the perspective of probability theory in order to train optimal classifiers. Specifically, we cast our inputs X and predictions Y as **random variables**, and define **joint probability distributions** that capture the interdependence of X and Y. If none of this makes sense right now, don't worry–we'll cover each of these terms in detail. But one important point to note is that classifiers trained within this framework do not require the input data to be linearly separable. This makes them quite powerful and capable of handling larger, more complex datasets than the simple perceptron or hard margin SVM classifiers that we have studied thus far.

To motivate Bayes decision theory, let's begin by presenting an example. Suppose that we are operating a medical clinic, and we want to classify whether a given patient that has come into our clinic has cancer ($Y = y_1$) or not ($Y = y_0$). In order to do this, we ask the patient how many calories they consume on a daily basis. This scalar value will be our input X for each patient. Let's use decision theory to predict whether a patient has cancer given their caloric intake.

2 Background

2.1 Bayes Theorem

Bayes theorem presents a simple relationship describing the probability of an event occurring, given that another event has already occurred. Let X and Y denote random variables that describe 2 different events. Then, Bayes theorem is as follows:

$$P(Y = y|X = x) = \frac{P(X = x \cap Y = y)}{P(X = x)} = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)} \tag{1}$$

This relationship allows us to write the probability of Y given X in terms of the probability of X given Y, which can be truly powerful as we will see shortly. Here is some terminology that we commonly use to denote these terms:

Prior probability: $P(Y = y)$, the probability of observing the outcome $Y = y$ "prior" to observing anything about X.

Generation probability/likelihood: $P(X = x|Y = y)$, the probability of observing the outcome $X = x$ given that $Y = y$, i.e. the probability of $Y = y$ "generating" the observed event $X = x$. This is also known as the **likelihood**.

Posterior probability: $P(Y = y|X = x)$, the probability that $Y = y$ given the observed outcome of $X = x$, i.e. the distribution of Y after observing everything about X.

Note: the prior or posterior probabilities for all outcomes $Y = y$ should always sum to 1.

2.2 Total Probability

In the Bayes equation above, we note the denominator term $P(X = x)$. We can evaluate this term using the total probability theorem on Y. Suppose that Y only has 2 outcomes, y_0 and y_1. Then, the total probability partition of X on Y is as follows:

$$P(X = x) = P(X = x|Y = y_0)P(Y = y_0) + P(X = x|Y = y_1)P(Y = y_1) \tag{2}$$

This rule generalizes well to cases where Y has n outcomes: simply partition on each outcome.

3 Risk Minimization

Let's get back to our cancer example. Before we can move forward, we need to know what penalty we incur when we make an incorrect prediction, i.e. predict that patient 1 has cancer (y_1) when in fact they do not (y_0). For now, let's assume a **0-1 loss function**, where we incur a penalty of 1 if we are wrong and 0 if we are right. Say y_p = our prediction and y_a = actual class:

$$L(y_p,\ y_a) = \begin{cases} 0 & y_p = y_a \\ 1 & y_p \neq y_a \end{cases} \tag{3}$$

Now, we have to train our classifier. In this setting, **training our classifier with training data (X, Y) means that we want to use this training data to estimate our prior distributions $P(Y = y)$ and likelihoods $P(X = x|Y = y)$**. Suppose that we look at our clinic's logs for the past month, and see the following statistics for 100 no-cancer and 100 cancer patients:

Daily Caloric Intake (X)	No-Cancer Patients (y_0)	Cancer Patients (y_1)
< 1200	1	20
$1200 - 1600$	10	50
> 1600	89	30

Suppose we also know that, in the general population, 10% of people have cancer and 90% of people do not. These are our priors, i.e. $P(Y = y_0) = 0.9$ and $P(Y = y_1) = 0.1$. If we did not have these general population numbers, we could have simply estimated our priors using the total number of cancer/no-cancer patients we have seen in the past.

From this training data, we can estimate the likelihoods $P(X = x|Y = y)$. For simplicity, let's assume that X is a discrete random variable with 3 caloric "buckets": < 1200, $1200 - 1600$, and > 1600 calories. Then, using the numbers directly from the training data, we might calculate $P(X = 1200 - 1600|Y = y_0) = 0.1$ and $P(X = 1200 - 1600|Y = y_1) = 0.5$, and so on. This is ideal: we now have estimates of our priors and likelihoods, enabling us to apply Bayes theorem.

Suppose that a patient walks into our clinic and tells us that he consumes $X = 1400$ calories per day. Since we have our priors/likelihoods, **we can use Bayes theorem to compute the posterior probabilities** $P(Y = y_0|X = 1200 - 1600)$ **and** $P(Y = y_1|X = 1200 - 1600)$. Doing the math, we see $P(Y = y_0|X = 1200 - 1600) = 0.64$ and $P(Y = y_1|X = 1200 - 1600) = 0.36$. Just by knowing the patient's caloric intake, we estimate that their propensity for cancer – 36% – is substantially higher than that of a random person from the general population (10%).

However, which class should we actually predict for this patient? Here is where our 0-1 loss comes in. The **risk of our classification** is defined as the expected value of the loss we might incur:

$$\begin{aligned} R \text{ (risk)} = E[L(y_p, y_a)] &= E[L(y_p, y_a = y_0)]P(Y = y_0|X) + E[L(y_p, y_a = y_1)]P(Y = y_1|X) \\ &= L(y_p, y_0)P(Y = y_0|X) + L(y_p, y_1)P(Y = y_1|X) \end{aligned} \tag{4}$$

The first equality follows by total expectation over the partitioning events $P(Y = y|X)$, and the second follows from the fact that, given a specific y_p, the loss is a constant, hence the expected value of the loss is the loss itself. Suppose we predict $y_p = y_0$, no-cancer. Then, the risk becomes $0 * 0.64 + 1 * 0.36 = 0.36$. If we instead predict $y_p = y_1$, that the patient has cancer, then the risk becomes $1 * 0.64 + 0 * 0.36 = 0.64$. This leads us to a key point: **for the 0-1 loss, the risk of predicting a certain class is precisely the posterior probability of the other class**.

Any good classifier would seek to **minimize the risk** of its prediction. Therefore, **for the 0-1 loss, in order to minimize risk, we always predict the class with the higher posterior probability**. Intuitively, the risk then would be the lower posterior probability; we have thus minimized risk. Therefore, for this patient, we predict y_0, that the patient does not have cancer.

4 Bayes Optimal Classifier

We can generalize the analysis of the previous section to devise a **decision rule** that the hospital can follow for any new patient that enters the clinic. Recall that a decision rule $r(x) \to y$ will give us a prediction y_p for any input x. We can calculate all of the posteriors in the following table:

Daily Caloric Intake (X)	No-Cancer Posteriors (y_0)	Cancer Posteriors (y_1)
< 1200	0.31	0.69
$1200 - 1600$	0.64	0.36
> 1600	0.96	0.04

Based on this posterior table (and assuming 0-1 loss), the **risk-minimizing decision rule**, also known as the **Bayes optimal decision rule**, is simply the decision rule that predicts the higher posterior for each observed value of X. In this case, we predict y_1 if $x < 1200$, else y_0. We specify our **decision boundary** to be $x = 1200$ calories, since this demarcates which class we predict.

5 Alternate Loss Functions

Thus far, we have assumed the relatively simplistic 0-1 loss setting (equation 3). However, for our cancer detection scenario, consider the true impact of **false positives** (predict y_1 when patient doesn't have cancer) vs **false negatives** (predict y_0 when patient has cancer). For false positives, we might refer the patient for additional testing, such as a C-T scan, to more accurately diagnose their condition; at worst, we would incur the cost of this additional operation, only to find out that the patient does not have cancer. However, for the false negative, we would turn the patient away thinking that they are cancer-free; they would live their life as before, their condition would deteriorate over time, and they might die because of our negligence. Therefore, we would like to heavily penalize false negatives relative to false positives; intuitively, we want to err on the side of caution and predict y_1 even if there is a slim chance that the patient has cancer.

To do this within our Bayes optimal classifier setting, we can augment our loss function. Suppose we have the new loss function below, penalizing false negatives 5X more than false positives:

$$L(y_p,\ y_a) = \begin{cases} 0 & y_p = y_a \\ 1 & y_p = y_1 \text{ and } y_a = y_0 \text{ (false positive)} \\ 5 & y_p = y_0 \text{ and } y_a = y_1 \text{ (false negative)} \end{cases} \tag{5}$$

To devise a decision rule $r(x)$ for the new loss setting, we again want to enumerate and minimize risk $R(r)$. This time, our risk derivation is a bit more complicated, as we are incorporating X into our model as well (since we want to generalize our decision rule to minimize risk for all $X = x$, we have to incorporate the posteriors $P(Y = y|X = x)$ for all $x,\ y$ into our expected loss calculation).

$$\begin{aligned} R(r) = E[L(y_p, y_a)] &= \sum_{\forall x} \Big[E[L(y_p, y_a)|X = x] * P(X = x) \Big] \quad \text{(total expectation)} \\ &= \sum_{\forall x} \bigg[\Big(E[L(y_p, y_0)] * P(Y = y_0|X = x) + E[L(y_p, y_1)] * P(Y = y_1|X = x) \Big) * P(X = x) \bigg] \\ &= \sum_{\forall x} \bigg[\Big(L(y_p, y_0)P(Y = y_0|X = x) + L(y_p, y_1)P(Y = y_1|X = x) \Big) * P(X = x) \bigg] \end{aligned} \tag{6}$$

The second line above again follows from total expectation (this time on $Y = y$), and the third line follows from the fact that the loss with fixed y_a is constant, hence its expectation is itself. For our cancer example, X only has 3 values x (the 3 calorie levels), so we would simply sum over these values and calculate posteriors/losses. We can alternatively first perform total expectation on the events $Y = y$ (rather than $X = x$), to get the below equivalent formulation of risk:

$$\begin{aligned} R(r) = E[L(y_p, y_a)] &= \sum_{\forall y} \Big[E[L(y_p, y)|Y = y] * P(Y = y) \Big] \quad \text{(total expectation)} \\ &= P(Y = y_0) * E[L(y_p, y_0)|Y = y_0] + P(Y = y_1) * E[L(y_p, y_1)|Y = y_1] \\ &= P(Y = y_0) \sum_{\forall x} \Big[E[L(y_p, y_0)] * P(X = x|Y = y_0) \Big] \\ &\qquad + P(Y = y_1) \sum_{\forall x} \Big[E[L(y_p, y_1)] * P(X = x|Y = y_1) \Big] \\ &= P(Y = y_0) \sum_{\forall x} \Big[L(y_p, y_0)P(X = x|Y = y_0) \Big] + P(Y = y_1) \sum_{\forall x} \Big[L(y_p, y_1)P(X = x|Y = y_1) \Big] \end{aligned} \tag{7}$$

The second line comes from enumerating the $Y = y$ cases, the third line from total expectation on the $X = x$ events, and the final line from taking expectation of a constant loss value.

These 2 formulations of risk are equivalent. The later might be easier to compute if you have the likelihoods and priors, while the former would be easier to compute if you have the posteriors and the probabilities $P(X = x)$ for each x. For our cancer example, we would have to compute each $P(X = x)$ using total probability – a cumbersome task – so we opt for the later formulation.

What decision rule $r(x_o)$ (x_o = observed x) minimizes risk? Taking the former definition of risk, consider how a single datapoint $X = x_o$ contributes to total risk. This **risk contribution** for x_o can be expressed as follows, by simply isolating the $X = x_o$ term from the $R(r)$ summation:

$$R(r, x_o) = \big[L(y_p, y_0)P(Y = y_0|X = x_o) + L(y_p, y_1)P(Y = y_1|X = x_o)\big] * P(X = x_o) \tag{8}$$

Now, **minimizing total risk is equivalent to minimizing the risk contributions for each $X = x_o$ in our training set**. To minimize the risk contribution for a particular x_o, note that the $P(X = x_o)$ term does not matter, as we have no control over it: it will be the same regardless of our prediction $r(x_o) = y_p$. Thus, we can ignore this term and strive to minimize the inner sum:

$$r(x_o) = \arg\min_{y_p} L(y_p, y_0)P(Y = y_0|X = x_o) + L(y_p, y_1)P(Y = y_1|X = x_o) \tag{9}$$

This formulation looks familiar: it is analogous to equation 4, except we now have an asymmetric loss function and have cast the problem as an optimization task. To solve this optimization, we must choose the class y_p corresponding to the larger **loss-weighted posterior probability** term out of $L(y_p, y_0)P(Y = y_0|X = x_o)$ and $L(y_p, y_1)P(Y = y_1|X = x_o)$. Doing so would zero-out the larger term and leave us with the smaller term as the risk contribution for x_o. Here is the rule:

$$r(x_o) = \begin{cases} y_0 & L(y_1, y_0)P(Y = y_0|X = x_o) > L(y_0, y_1)P(Y = y_1|X = x_o) \\ y_1 & \text{otherwise} \end{cases} \tag{10}$$

This decision rule minimizes our risk formulation, because by predicting the class with the higher loss-weighted posterior, our risk becomes the lower loss-weighted posterior, which by definition is minimum. For our cancer example, we have the below loss-weighted (LW) posterior probabilities:

Daily Caloric Intake (X)	No-Cancer LW-Posteriors (y_0)	Cancer LW-Posteriors (y_1)
< 1200	0.31*1 = 0.31	0.69*5 = 3.45
$1200 - 1600$	0.64*1 = 0.64	0.36*5 = 1.80
> 1600	0.96*1 = 0.96	0.04*5 = 0.20

Therefore, for our cancer example with augmented loss function, our decision rule becomes $r(x) = y_1$ if $x < 1600$, else predict y_0. Our decision boundary has become $x = 1600$ calories. This makes sense intuitively: since false negatives are penalized heavily, we have a higher threshold (1600 vs 1200 calories) that a patient has to meet in terms of their caloric intake before they can be safely classified as not having cancer. We err on the side of caution, predicting y_1 more often.

As a final note, all of this analysis extends analogously to the setting where X is a continuous (rather than discrete) random variable. Our likelihoods become continuous distributions, and we would have to estimate these distributions via more rigorous methods (simple counting no longer works). This is the intuition behind Gaussian discriminant analysis (LDA/QDA): we model the likelihood distributions using Gaussians estimated from the training data (fitting sample mean and variance), and use the same formula for risk to arrive at the same Bayes optimal decision rule.