ARTIFICIAL INTELLIGENCE

Assignment 2

Submitted by
Adil Hussain Mughal (12084)

BS (SE-5th) MORNING

Date: 13 June 2021

Submitted to: Sir Hammad Dilpazir

DEPARTMENT OF ENGINEERING NATIONAL UNIVERSTIY OF MODERN LANGUAGES, ISLAMABAD

Question

Write a note on Bayesian Decision Theory. Your answer should include (but not limited to) the following questions.

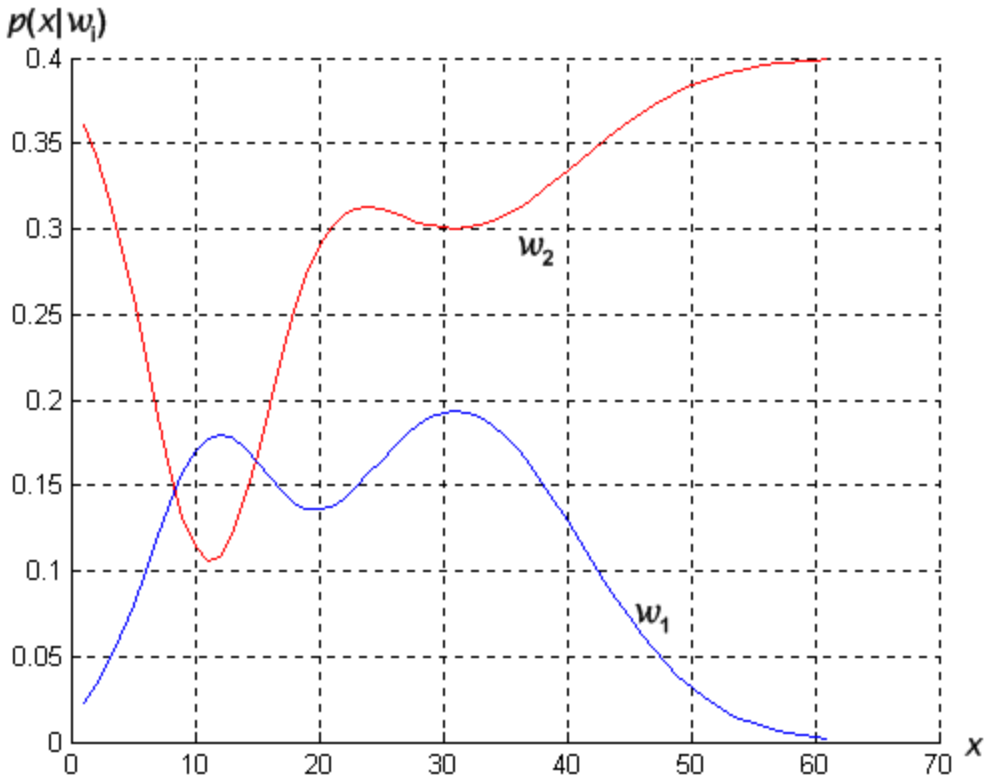
- **How do you we find the optimal decision boundary?**
- **What does Bayes decision theory optimize when making decisions?**
- **What are the tradeoffs in Bayes decision theory**?

Bayesian Decision Theory

Bayesian Decision Theory is a fundamental statistical approach to the problem of pattern classification. Quantifies the tradeoffs between various classifications using probability and the costs that accompany such classifications. Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. It is considered the ideal case in which the probability structure underlying the categories is known perfectly. While this sort of stiuation rarely occurs in practice, it permits us to determine the optimal (Bayes) classifier against which we can compare all other classifiers. Moreover, in some problems it enables us to predict the error we will get when we generalize to novel patterns.

This approach is based on quantifying the tradeoffs between various classification decisions using probability and the costs that accompany such decisions. It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known.

Let us reconsider the hypothetical problem posed in Chapter 1 of designing a classifier to separate two kinds of fish: sea bass and salmon. Suppose that an observer watching fish arrive along the conveyor belt finds it hard to predict what type will emerge next and that the sequence of types of fish appears to be random. In decision-theoretic terminology we would say that as each fish emerges nature is in one or the other of the two possible states: Either the fish is a sea bass or the fish is a salmon. We let w denote the state of nature, with w = w1 for sea bass and w = w2 for salmon. Because the state of nature is so unpredictable, we consider w to be a variable that must be described probabilistically.



If the catch produced as much sea bass as salmon, we would say that the next fish is equally likely to be sea bass or salmon. More generally, we assume that there is some prior probability P(w1) that the next fish is sea bass, and some prior probability P(w2) that it is salmon. If we assume there are no other types of fish relevant here, then P(w1)+ P(w2)=1. These prior probabilities reflect our prior knowledge of how likely we are to get a sea bass or salmon before the fish actually appears.

If we are forced to make a decision about the type of fish that will appear next just by using the value of the prior probabilities we will decide w1 if P(w1)> P(w2) otherwise decide w2. This rule makes sense if we are to judge just one fish, but if we were to judge many fish, using this rule repeatedly, we would always make the same decision even though we know that both types of fish will appear. Thus, it does not work well depending upon the values of the prior probabilities.

In most circumstances, we are not asked to make decisions with so little information. We might for instance use a lightness measurement x to improve our classifier. Different fish will yield different lightness readings, and we express this variability: we consider x to be a continuous random variable whose distribution depends on the state of nature and is expressed as p(x|w). This is the class-conditional probability density (state-conditional probability density) function, the probability density function for x given that the state of nature is in w. Then the difference between p(x|w1) and p(x|w2) describes the difference in lightness between populations of sea bass and salmon

Suppose that we know both the prior probabilities P(wj) and the conditional densities p(x|wj) for j = 1, 2. Suppose further that we measure the lightness of a fish and discover that its value is x. How does this measurement influence our attitude concerning the true state of nature? We note first that the (joint) probability density of finding a pattern that is in category wj and has feature value x can be written in two ways: p(wj,x) = P(wj|x) p(x) = p(x|wj) P(wj). Rearranging these leads us to the answer to our question, which is called Bayes formula:

$$P(w_j \mid x) = \frac{p(x \mid w_j) P(w_j)}{p(x)}$$

where in this case of two categories

$$p(x) = \sum_{j=1}^{2} p(x \mid w_j) P(w_j)$$

Bayes formula can be expressed informally as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

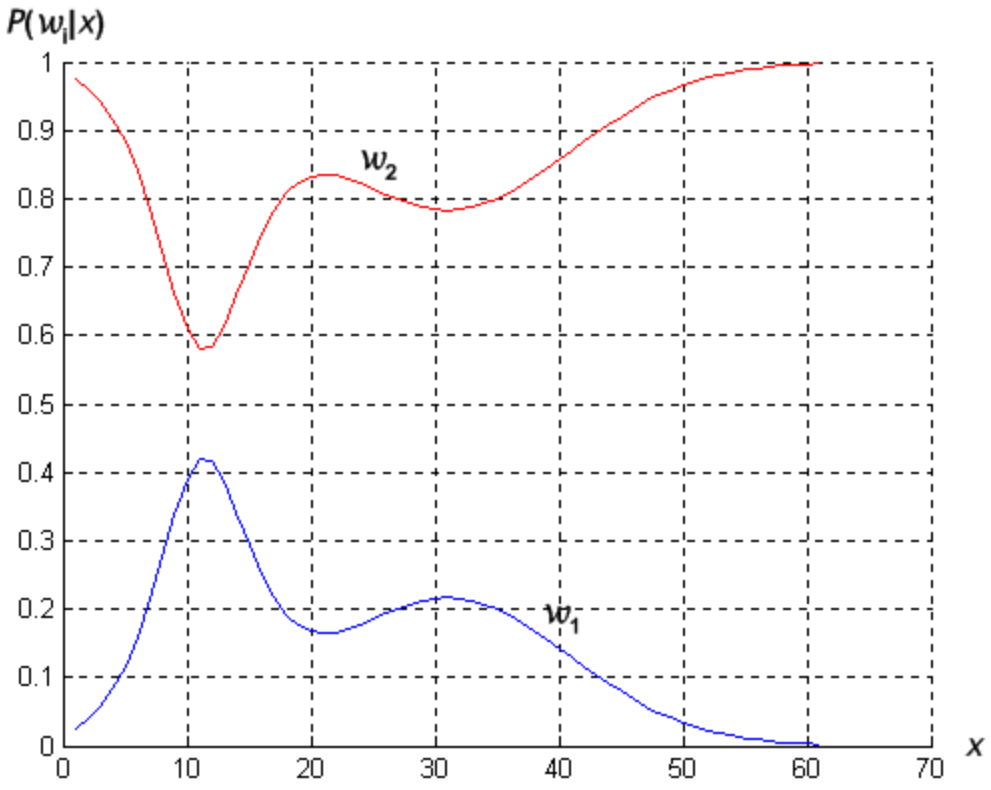
Bayes formula shows that by observing the value of x we can convert the prior probability P(wj) to the posterior probability P(wj|x) -the probability of the state of nature being wj given that feature value x has been measured. p(x|wj) is called the likelihood of wj with respect to x, a term chosen to indicate that, other things being equal, the category wj, for which p(x|wj) is large is more “likely” to be the true category. Notice that it is the product of the likelihood and the prior probability that is most important in determining the posterior probability; the evidence factor p(x), can be viewed as a scale factor that guarantees that the posterior probabilities sum to one. The variation of posterior probability P(wj|x) with x is illustrated in Figure 4.2 for the case P(w1)=2/3 and P(w2)=1/3.

If we have an observation x for which P(w1|x)>P(w2|x), we would naturally be inclined to decide that the true state of nature is w1. The probability of error is calculated as

$$P(error \mid x) = \begin{cases} P(w_1 \mid x) & \text{if we decide } w_2 \\ P(w_2 \mid x) & \text{if we decide } w_1 \end{cases}$$

The Bayes decision rule is stated as

Decide w1 if P(w1|x)>P(w2|x); otherwise decide w2



P(error|x)=min[P(w1|x), P(w2|x)]

This form of decision rule emphasizes the role of the posterior probabilities. As being equivalent, the same rule can be expressed in terms of conditional and prior probabilities as:

Decide w1 if p(x|w1)P(w1) > p(x|w2)P(w2); otherwise decide w2

Bayesian Decision Theory

We shall now formalize the ideas just considered, and generalize them in four ways: by allowing the use of more than one feature, by allowing more than two states of nature, by allowing actions other than merely deciding the state of nature, and by introducing a loss function more general than the probability of error. Allowing the use of more than one feature merely requires replacing the scalar x by the feature vector x, where x is in a d-dimensional Euclidean space Rd called the feature space. Allowing more than two states of nature provides us with a useful

generalization for a small notational expense as {w1… wc}. Allowing actions other than classification as {a1…aa} allows the possibility of rejection-that is, of refusing to make a decision in close (costly) cases. The loss function states exactly how costly each action is, and is used to convert a probability determination into a decision. Cost functions let us treat situations in which some kinds of classification mistakes are more costly than others. Then the posterior probability can be computed by Bayes formula as:

$$P(w_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid w_j)P(w_j)}{p(\mathbf{x})}$$

where the evidence is now

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x} \mid w_j)P(w_j)$$

Suppose that we observe a particular x and that we contemplate taking action ai. If the true state of nature is wj by definition, we will incur the loss l(ai|wj). Because P(wj|x) is the probability that the true state of nature is wj, the expected loss associated with taking action ai is

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \mathbf{x})P(w_j \mid \mathbf{x})$$

An expected loss is called a risk, and R(ai|x) is called the conditional risk. Whenever we encounter a particular observation x, we can minimize our expected loss by selecting the action that minimizes the conditional risk.

If a general decision rule a(x) tells us which action to take for every possible observation x, the overall risk R is given by

$$R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x})p(\mathbf{x})d\mathbf{x}$$

Thus, the Bayes decision rule states that to minimize the overall risk, compute the conditional risk given in Eq.4.10 for i=1…a and then select the action ai for which R(ai|x) is minimum. The resulting minimum overall risk is called the Bayes risk, denoted R, and is the best performance that can be achieved.

Two-Category Classification

When these results are applied to the special case of two-category classification problems, action a1 corresponds to deciding that the true state of nature is w1, and action a2 corresponds to deciding that it is w2. For notational simplicity, let lij=l(ai|wj) be the loss incurred for deciding wi, when the true state of nature is wj. If we write out the conditional risk given by Eq.4.10, we obtain

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11}P(w_1 \mid \mathbf{x}) + \lambda_{12}P(w_2 \mid \mathbf{x})$$

$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21}P(w_1 \mid \mathbf{x}) + \lambda_{22}P(w_2 \mid \mathbf{x})$$

There are a variety of ways of expressing the minimum-risk decision rule, each having its own minor advantages. The fundamental rule is to decide w1 if R(a1|x)<R(a2|x). In terms of the posterior probabilities, we decide w1 if

R(a1|x)<R(a2|x)

$$\lambda_{11}P(w_1 \mid \mathbf{x}) + \lambda_{12}P(w_2 \mid \mathbf{x}) < \lambda_{21}P(w_1 \mid \mathbf{x}) + \lambda_{22}P(w_2 \mid \mathbf{x})$$

$$(\lambda_{21} - \lambda_{11})P(w_1 \mid \mathbf{x}) > (\lambda_{12} - \lambda_{22})P(w_2 \mid \mathbf{x})$$

or in terms of the prior probabilities

$$(\lambda_{21} - \lambda_{11})p(\mathbf{x} \mid w_1)P(w_1) > (\lambda_{12} - \lambda_{12})p(\mathbf{x} \mid w_2)P(w_2)$$

or alternatively as likelihood ratio

$$\frac{p(\mathbf{x} \mid w_1)}{p(\mathbf{x} \mid w_2)} > \frac{(\lambda_{12} - \lambda_{12})}{(\lambda_{21} - \lambda_{11})} \frac{P(w_2)}{P(w_1)}$$

This form of the decision rule focuses on the x-dependence of the probability densities. We can consider p(x|wj) a function of wj (i.e., the likelihood function) and then form the likelihood ratio p(x|w1)/ p(x|w2). Thus the Bayes decision rule can be interpreted as calling for deciding w1 if the likelihood ratio exceeds a threshold value that is independent of the observation x.

Minimum Error Rate Classification

In classification problems, each state of nature is associated with a different one of the classes, and the action ai is usually interpreted as the decision that the true state of nature is wi. If action ai is taken and the true state of nature is wj then the decision is correct if i=j and in error if i¹j. If errors are to be avoided it is natural to seek a decision rule, that minimizes the probability of error, that is the error rate.

This loss function is so called symetrical or zero-one loss function is given as

$$\lambda(\alpha_i \mid \mathbf{x}) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

This loss function assigns no loss to a correct decision, and assigns a unit loss to any error: thus, all errors are equally costly. The risk corresponding to this loss function is precisely the average probability of error because the conditional risk for the two-category classification is

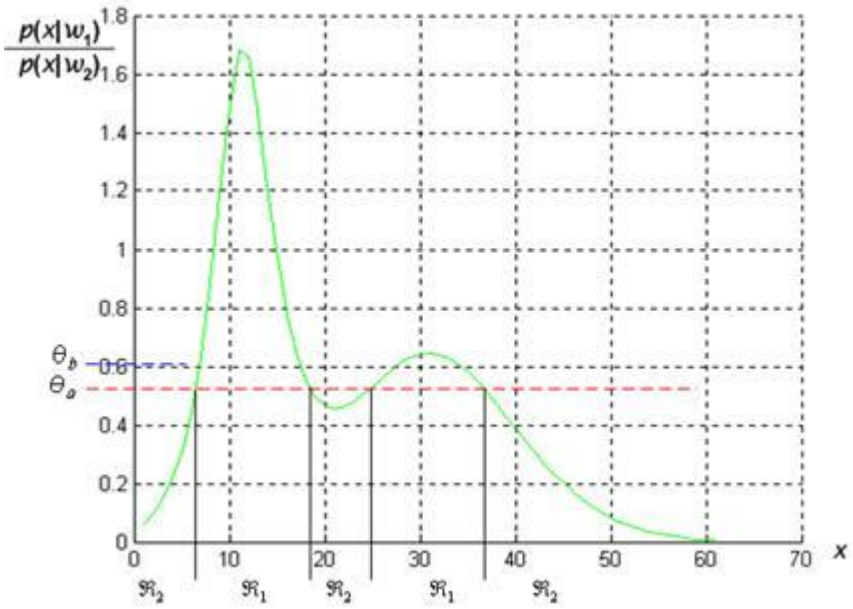
$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \mathbf{x})P(w_j \mid \mathbf{x})$$

$$= \sum_{j \neq i}^{c} P(w_j \mid \mathbf{x})$$

$$= 1 - P(w_j \mid \mathbf{x})$$

and P(wj|x) is the conditional probability that action ai is correct. The Bayes decision rule to minimize risk calls for selecting the action that minimizes the conditional risk. Thus, to minimize the average probability of error, we should select the i that maximizes the posterior probability P(wj|x). In other words, for minimum error rate:

Decide wi if P(wi|x)>P(wj|x) for all i[1]



Bayesian decision theory refers to a decision theory which is informed by Bayesian probability. It is a statistical system that tries to quantify the tradeoff between various decisions, making use of probabilities and costs. An agent operating under such a decision theory uses the concepts of Bayesian statistics to estimate the expected value of its actions, and update its expectations based on new information. These agents can and are usually referred to as estimators.

From the perspective of Bayesian decision theory, any kind of probability distribution such as the distribution for tomorrow's weather represents a prior distribution. That is, it represents how we expect today the weather is going to be tomorrow. This contrasts with frequentist inference, the classical probability interpretation, where conclusions about an experiment are drawn from a set of repetitions of such experience, each producing statistically independent results. For a frequentist, a probability function would be a simple distribution function with no special meaning.

Suppose we intend to meet a friend tomorrow, and expect an 0.5 chance of raining. If we are choosing between various options for the meeting, with the pleasantness of some of the options (such as going to the park) being affected by the possibility of rain, we can assign values to the different options with or without rain. We can then pick the option whose expected value is the highest, given the probability of rain.

- One definition of rationality, used both on Less Wrong and in economics and psychology, is behavior which obeys the rules of Bayesian decision theory. Due to computational constraints, this is impossible to do perfectly, but naturally evolved brains do seem to mirror these probabilistic methods when they adapt to an uncertain environment. Such models and distributions may be reconfigured according to feedback from the environment.

How do you we find the optimal decision boundary?

A decision boundary is a graphical representation of the solution to a classification problem. Decision boundaries can help us to understand what kind of solution might be appropriate for a problem. They can also help us to understand the how various machine learning classifiers arrive at a solution. The optimal decision boundary represents the "best" solution possible for that problem. Consequently, by looking at the complexity of this boundary and at how much error it produces, we can get an idea of the inherent difficulty of the problem. Unless we have generated the data ourselves, we won't usually be able to find the optimal boundary. Instead, we approximate it using a classifier. A good machine learning classifier tries to approximate the optimal boundary for a problem as closely as possible.

Optimal Boundaries

A classification problem asks: given some observations of a thing, what is the best way to assign that thing to a class based on some of its features? For instance, we might want to predict whether a person will like a movie or not based on some data we have about them, the "features" of that person.

A solution to the classification problem is a rule that partitions the features and assigns each all the features of a partition to the same class. The "boundary" of this partitioning is the **decision boundary** of the rule.

It might be that two observations have exactly the same features, but are assigned to different classes. (Two things that look the same in the ways we've observed might differ in ways we haven't observed.) In terms of probabilities this means both

$$P(C=0|X)>0P(C=0|X)>0$$

and

$$P(C=1|X)>0P(C=1|X)>0$$

In other words, we might not be able with full certainty to classify an observation. We could however assign the observation to its *most probable* class. This gives us the decision rule

$$C^=argmaxcP(C=c|X)C^=argmaxc P(C=c|X)$$

The boundary that this rule produces is the **optimal decision boundary**. It is the MAP estimate of the class label, and it is the rule that minimizes classification error under the zero-one loss function. We will look at error and loss more in a future post.

We will consider *binary* classification problems, meaning, there will only be two possible classes, 0 or 1. For a binary classification problem, the optimal boundary occurs at those points where each class is equally probable:

$$P(C=0|X)=P(C=1|X)P(C=0|X)=P(C=1|X)$$

What does Bayes decision theory optimize when making decisions?

Bayesian decision making involves basing decisions on the probability of a successful outcome, where this probability is informed by both prior information and new evidence the decision maker obtains. The statistical analysis that underlies the calculation of these probabilities is Bayesian analysis. In recent years, the Bayesian approach has been applied more commonly in both nutrition research and clinical decision making, and registered dietitian nutritionists would benefit from gaining a deeper understanding of this approach. This article provides a background of Bayesian decision making and analysis, and then presents applications of the approach in two different areas-medical diagnoses and nutrition policy research. It concludes with a description of how Bayesian decision making may be used in everyday life to allow each of us to appropriately weigh established beliefs and prior knowledge with new data and information in order to make well-informed and wise decisions.

Conclusion

What you have just learned is a simple, univariate application of Bayesian Decision Theory that can be expanded onto a larger feature space by using the multivariate Gaussian distribution in place of the evidence and likelihood. Although this article focused on tackling the problem of cancer detection, Bayes' Theorem is used in a variety of disciples including investing, marketing, and systems engineering.

https://github.com/Enggadil/-AI-LAB-_BSSE-5-M-