Essay 12: Bayesian Decision Theory

Raul Adell

May 28, 2023

1 Introduction

This essay draws upon the book "Pattern Classification" by Richard O. Duda, Peter E. Hart, and David G. Stork as the main reference. It explores the fundamental concepts of Bayesian decision theory, a mathematical framework that combines probability theory and decision theory to tackle decision-making under uncertainty.

Key topics covered include the role of prior and posterior distributions, which capture beliefs about unknown parameters or hypotheses, and the importance of informative priors and their update using Bayes' theorem. Likelihood functions and their interpretation in data consistency and parameter estimation are also discussed.

The essay delves into decision rules within the Bayesian framework, focusing on discriminant functions for class assignment, and explores the Minimum-Error-Rate classification approach to minimize misclassification probabilities. The Two-Category Case is studied as a foundational example with broader applications.

2 Main concepts

In Bayesian decision theory, understanding the concepts of state of nature, **prior** distribution, and decision rule is crucial for making optimal decisions under uncertainty. To illustrate these concepts, let's consider an example of a medical diagnosis problem.

Imagine a scenario where a patient comes to a clinic with certain symptoms, and the goal is to determine whether they have a particular disease (let's call it Disease X) or not. In this context, we can define the **state of nature** as the true condition of the patient, which can be either having Disease (state w_1) or not having Disease X (state w_2). Thus the total amount of states is $w = \{w_1, w_2\}$.

The prior distribution captures our beliefs before considering any specific information about the patient's symptoms. It reflects our initial understanding of the likelihood of each state of nature based on available data or expert opinions.

Suppose we are forced to make a decision about the state of a patient that has just come in. Assume also (not realistically) that any classification entrails the same cost or consequences. If a decision has to be made with so little information it makes sense to make use of the prior probabilities. Thus the following **decision rule** seems rather logical: Decide w_1 if $P(w_1) > P(w_2)$; otherwise decide w_2.

This way, the **probability error** is the smaller of $P(w_1)$ and $P(w_2)$. We shall see later that no other decision rule can yield a larger probability of being right.

Consider now that we gain a little bit more knowledge on our patient's blood pressure x. x could be any other feature related with the patient. We consider x to be a continuous random variable whose distribution depends o the state of nature, and is expressed as . This is the **class-conditional**

probability density function. Then the difference between $p(x, w_1)$ and $p(x, w_2)$ describes the difference in blood pressure between classes of patients.

Suppose now that we measure x. How does this measurement influence our attitude concerning the state of nature of the patient? We note first that the (joint) probability density of finding a pattern that is in category w_1 and has feature value x is: $p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j)$, where only the multiplication probability rule was used $P(A \cap B) = P(A|B) * P(B)$. Rearranging this leads to **Bayes' formula**:

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},$$

where in this case with only two categories, one can use the total probability formula to express the denominator as:

$$p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j).$$

Another more informal way to look at Bayes' formula would be as a way to update the knowledge (**posterior**) by incorporating new information of x:

$$posterior = \frac{likelihood \ * \ prior}{evidence}.$$

We call $p(x|\omega_j) = L(\omega_j|x)$ the **likelihood** of ω_j with respect to x. That is, how well ω_j value explains the observed data x. Notice the product of likelihood and the prior probability is key in order to update the current knowledge or posteriori probablity. However, the evidence is just a normalization factor.

Given that the knowledge on a feature x has been incorporated in the a posteriori probability, we would like to decide the natural state of nature while trying minimize the prediction error. A good decision rule would be to choose w_1 if $P(w_1|x) > P(w_2|x)$; choose w_2 otherwise.

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

Clearly, for a given x we can minimize the probability if error by applying the previously mentioned decision rule. In fact this rule minimizes the average probability of error:

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x)\, dx = \int_{-\infty}^{\infty} P(\text{error}|x)p(x)\, dx.$$

By deciding w_1 if $P(w_1|x) > P(w_2|x)$; choosing w_2 otherwise, we ensure that for every x, $P(\text{error}|x)$ is as small as possible. The previous decision rule is called **Bayes' decision rule** and minimizes the probability of error.

In the same way the minimum of the priori probabilities defined the error of prediction without information, now the minimum posteriori defines the error of prediction taking into account information in x:

$$P(\text{error}|x) = \min\{P(w_1|x), P(w_2|x)\}$$

MAP or maximum a posteriori estimation can be seen as the application of Bayes' decision rule specifically to the problem of parameter estimation. In MAP estimation, the goal is to find the parameter value that maximizes the posterior probability given the observed data and prior knowledge.

3 Formalization

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
- 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 $\mathbf{x}$, where $\mathbf{x}$ is in a d-dimensional Euclidean space R^d, called the **feature space**.

Formally, 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.

Let $\omega_1, \ldots, \omega_c$ be the finite set of c states of nature ("categories") and $\alpha_1, \ldots, \alpha_a$ be the finite set of a possible actions. The loss function $\lambda(\alpha_i|\omega_j)$ describes the loss incurred for taking action α_i when the state of nature is ω_j. Let the feature vector $\mathbf{x}$ be a d-component vector-valued random variable, and let $p(\mathbf{x}|\omega_j)$ be the state-conditional probability density function for $\mathbf{x}$ - the probability density function for $\mathbf{x}$ conditioned on ω_j being the true state of nature.

The Bayes' formula has the same shape as in last section but now the evidence is:

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

Suppose that we observe a particular $\mathbf{x}$ and that we contemplate taking action α_i. If the true state of nature is ω_j, by definition we will incur the loss $\lambda(\alpha_i|\omega_j)$. The **expected loss or cost associated with taking action** α_i is merely:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}).$$

The **risk** is computed by summing the product of the loss function and the posterior probabilities over all possible states of nature ω_j. It takes into account both the potential losses for each possible state of nature and their respective probabilities. A lower risk implies a more desirable action in terms of cost. Whenever we encounter a particular observation $\mathbf{x}$, we can minimize our expected loss by selecting the action that minimizes the conditional risk. We shall now show that this Bayes decision procedure actually provides the optimal performance on an overall risk.

Stated formally, our problem is to find a decision rule against $P(\omega_j)$ that minimizes the overall risk. A **general decision rule** is a function $\alpha(\mathbf{x})$ that tells us which action to take for every possible observation $\mathbf{x}$.

$$R = \int R(\alpha(\mathbf{x}|\mathbf{x})p(\mathbf{x})\,dx$$

Clearly, if $\alpha(\mathbf{x})$ is chosen so that $R(\alpha(\mathbf{x}|\mathbf{x})$ is as small as possible for every $\mathbf{x}$, then the overall risk will be minimized.

This justifies the following statement of the Bayes decision rule: To minimize the overall risk, compute the conditional risk $R(\alpha_i|\mathbf{x})$ for all possible actions α_i from $i = 1, ..., i = a$ and select the action α_i for which $R(\alpha_i|\mathbf{x})$ is minimum. The resulting minimum overall risk when the true underlying distribution is known is called the **Bayes risk**, denoted R^*, and is the best performance that can be achieved.

3.1 Example: Two category classification

Here, action α_1 corresponds to deciding that the true state of nature is ω_1, and action α_2 corresponds to deciding that it is ω_2. For notational simplicity, let $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ be the loss incurred for deciding ω_i when the true state of nature is ω_j. If we write out the conditional risk, we obtain

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x})$$
$$R(\alpha_2|\mathbf{x}) = \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x}).$$

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 ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$. By using the previous equations we obtain:

$$(\lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x})$$

By expressing the posterior probability in terms of the likelihood and prior and also assuming that the loss incurred for making an error is greater than the loss incurred for being correct we get to:

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

This form of the decision rule focuses on the $\mathbf{x}$-dependence of the probability densities. By expressing it this way a new interpretation of the Bayes' decision rule arises. We decide ω_1 if the likelihood ration exceeds a given threshold value chosen by desired trade off between risks of different error types.

4 Minimum-Error-Rate Classification

Minimum-error-rate classification aims to minimize the probability of misclassification. The minimum-error-rate classification corresponds to the Bayes decision rule when the loss function assigns equal costs to different types of errors.

The action α_i is usually interpreted as the decision that the true state of nature is ω_i. If action α_i is taken and the true state of nature is ω_j, then the decision is correct if $i = j$, and in error if $i \neq j$. This cost function is called symmetrical or zero-one loss function and assumes all errors are equally costly. Other loss functions, such as quadratic and linear-difference find greater use in regression, for example.

If errors are to be avoided, it is natural to seek a decision rule that minimizes the probability of error, i.e., the error rate.

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}$$

The risk corresponding to this lost function is the probability of error:

$$\begin{aligned} R(\alpha_i|x) &= \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|x) \\ &= \sum_{j \neq i} P(\omega_j|x) \\ &= 1 - P(\omega_i|x) \end{aligned}$$

Bayes decision rule to minimize risk calls for selecting the action that maximizes the posterior probability $P(\omega_i|x)$. That is, the same rule we were familiar with.

Decide ω_i if $P(\omega_i|x) > P(\omega_j|x)$ for all $j \neq i$.

4.1 Minimax Criterion

The Minimax Criterion is a decision-making principle in statistical decision theory that aims to minimize the maximum possible risk or loss. It is a conservative approach that focuses on the worst-case scenario by considering the highest potential loss among all possible actions.

Mathematically, the Minimax Criterion can be formulated as follows: Given a set of actions or decisions α_i and a set of states of nature or outcomes ω_j, the goal is to find a decision rule that minimizes the maximum risk:

$$\min_{\alpha_i} \max_{\omega_j} R(\alpha_i|\omega_j)$$

This criterion seeks to find the decision rule that achieves the minimum value of the maximum risk over all possible actions and states of nature.

The Minimax Criterion is particularly useful in situations where the decision maker wants to be cautious and ensure robustness against the worst-case scenario. However, it does not take into account the probabilities or likelihoods of different outcomes and assumes a pessimistic viewpoint.

We again place ourselves in a two-category classification problem. In order to understand this, we let $\mathcal{R}_1$ denote the (as yet unknown) region in feature space where the classifier decides ω_1, and likewise, $\mathcal{R}_2$ represents the region where the classifier decides ω_2. By doing so, we can express our overall risk in terms of conditional risks.

$$\begin{aligned} R &= \int_{\mathcal{R}_1} R(\alpha_1|\mathbf{x})p(\mathbf{x}) + \int_{\mathcal{R}_2} R(\alpha_2|\mathbf{x})p(\mathbf{x})dx \\ &= \int_{\mathcal{R}_1} [\lambda_{11}P(\omega_1)p(\mathbf{x}|\omega_1) + \lambda_{12}P(\omega_2)p(\mathbf{x}|\omega_2)]dx + \int_{\mathcal{R}_2} [\lambda_{21}P(\omega_1)p(\mathbf{x}|\omega_1) + \lambda_{22}P(\omega_2)p(\mathbf{x}|\omega_2)]dx, \end{aligned}$$

where the Bayes' fromula $P(\omega_j|\mathbf{x}) = \frac{P(\omega_j)p(\mathbf{x}|\omega_j)}{p(\mathbf{x})}$ has been used. Using the fact that the possible state of nature has cardinality we can state:

- $P(\omega_2) = 1 - P(\omega_1)$
- $\int_{\mathcal{R}_1} f(\mathbf{x})dx = 1 - \int_{\mathcal{R}_2} f(\mathbf{x})dx$, where f is an arbitrary function.

Applying this to the previous equation we obtain:

$$\begin{aligned} R(P(\omega_1)) &= \overbrace{\lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x})|\omega_2)dx}^{R_{mm},\ \text{minimax risk}} \\ &+P(\omega_1) \underbrace{\left[(\lambda_{11} - \lambda_{22}) - (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)dx - (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)dx\right]}_{=0 \text{ for the minimax solution}} \end{aligned}$$

Thus, once the decision boundary is set ($\mathcal{R}_1$ and $\mathcal{R}_2$ determined), the overall risk is linear with $P(\omega_1)$. The equation also reveals that if we can found a boundary such that the constant of proportionality is 0, the risk is independent of the priors.

$$
\begin{aligned}
R_{\text{mm}} &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)\, dx \\
&= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \left(1 - \int_{R_2} p(\mathbf{x}|\omega_1)\, dx\right) \\
&= \lambda_{22} + (\lambda_{12} - \lambda_{22}) - (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)\, dx \\
&= \lambda_{12} + (\lambda_{22} - \lambda_{12}) \int_{\mathcal{R}_2} p(x|\omega_1)\, dx.
\end{aligned}
$$

Consider to be in the worst-case scenario, that is, prior probabilities make the Bayes risk maximum. This does not necessarily mean having no information about the priors. It means selecting the priors in a way that makes the decision rule perform poorly in terms of risk.

The way to proceed is by finding the the corresponding decision boundary (and therefore the decision rule) that provides the optimal solution that minimizes this case risk. Thus, error is mini

In game theory, you have a hostile opponent who can be expected to take an action maximally detrimental to you. Thus it makes great sense for you to take an action (e.g., make a classification) where your costs — due to your opponent's subsequent actions — are minimized.

For each value of the prior probabilities, there exists an optimal decision boundary and an associated Bayes error rate. When the priors are fixed and the decision boundary is determined, changing the prior probabilities will cause the probability of error to vary linearly with the value of $P(\omega 1)$. To minimize the maximum error, we should design our decision boundary based on the maximum Bayes error. By doing so, the error rate will remain constant and independent of the prior probabilities.

4.2 Neyman-Pearson Criterion

In some problems, we may wish to minimize the overall risk subject to a constraint; for instance, we might wish to minimize the total risk subject to the constraint $\int R(\alpha_i|x)dx <$ constant for some particular i. Such a constraint might arise when there is a fixed resource that accompanies one particular action α_i, or when we must not misclassify pattern from a particular state of nature ω_i at more than some limited frequency. For instance, in our fish example, there might be some government regulation that we must not misclassify more than 1 % of patients as healthy. We might then seek a decision that minimizes the chance of classifying a sea bass as a salmon subject to this condition. We generally satisfy such a Neyman-Pearson criterion by adjusting decision boundaries numerically. However, for Gaussian and some other distributions, Neyman-Pearson solutions can be found analytically.

5 Classifiers, Discriminant Functions and Decision Surfaces

There are many different ways to represent pattern classifiers. One of the most useful is in terms of a set of **discriminant functions** $g_i(x)$, $i = 1, \ldots, c$. The classifier is said to assign a feature vector x to class ω_i if

$$g_i(x) > g_j(x) \text{ for all } j \neq i.$$

Thus, the classifier is viewed as a network or machine that computes c discriminant functions and selects the category corresponding to the largest discriminant.

A Bayes classifier is easily and naturally represented in this way. For the general case with risks, we can let $g_i(x) = -R(\alpha_i|x)$, since the maximum discriminant function will then correspond to the minimum conditional risk. For the minimum-error-rate case, we can simplify things further by

taking $g_i(x) = P(\omega_i|x)$, so that the maximum discriminant function corresponds to the maximum posterior probability. Clearly, the choice of discriminant functions is not unique and is invariant upon transformations by a monotonically increasing function.

Even though the discriminant functions can be written in a variety of forms, the decision rules are equivalent. The effect of any decision rule is to divide the feature space into c decision regions, $\mathcal{R}_1, \ldots, \mathcal{R}_c$. The regions are separated by **decision boundaries**, surfaces in feature space where ties occur among the largest discriminant functions.

5.1 The Two-Category Case

Instead of using two discriminant functions g_1 and g_2 and assigning x to ω_1 if $g_1 > g_2$, it is more common to define a single discriminant function

$$g(x) \equiv g_1(x) - g_2(x).$$

Thus, a dichotomizer can be viewed as a machine that computes a single discriminant function $g(x)$ and classifies x according to the algebraic sign of the result.

6 Rest of the chapter

In the book the authors present the basic definitions of the Normal distribution. Those are later used in order to discuss different kinds of discriminant functions based over the normal distribution. In the last section of the decision theory chapter integral error probabilities bounds are discussed.

7 Conclusion

In conclusion, decision theory serves as a strong framework for machine learning due to its ability to handle uncertainty and provide principled decision-making. By incorporating uncertainty modeling, evaluation metrics, and contextual considerations, decision theory enhances the robustness, reliability, and ethical alignment of machine learning algorithms. It enables us to make informed choices, optimize performance based on various objectives, and address real-world complexities, leading to more effective and responsible machine learning systems.