Lecture Notes: Decision Theory for Classification

Ted Meeds[1,2]

[1] Informatics Institute, University of Amsterdam
[2] The Centre for Integrative Bioinformatics, Vrije University
tmeeds@gmail.com

Abstract In this note decision theory for classification is treated briefly. This is mainly from Bishop Chapter 1. Once a probabilistic model of classification data is built, the problem of making optimal decision remains. For regression, the optimal decision (for squared-loss) was the expected prediction. For classification, the optimal decision that minimizes misclassification rate is the assignment to the largest predicted class $P(C_k|\mathbf{x})$. However, there may be real-world costs for making mistakes, and for these cases the action taken will have the minimum expected loss under the $P(C_k|\mathbf{x})$ distributions.

1 Motivation for Decision Theory

During inference, fully probabilistic $p(C_k, \mathbf{x})$ or probabilistic discriminant $P(C_k|\mathbf{x})$ models are determined. Based on these models, decisions are made, but how do we make optimal decisions?

This note will focus on two aspects of decision theory for classification: 1) misclassification and 2) expected loss. For 1) recall we assigned class labels to the class with the largest probability; though intuitive, this note will describe why this is the optimal choice. For 2) consider real-world situations where making incorrect decisions have important consequences. As an example we will use a diagnosis test for cancer using an x-ray image. If the test predicts no cancer, but the patient actually has cancer, there is a very high cost.

2 Minimizing Misclassification Rate

Our intuition has been assign to class with highest $P(C_k|\mathbf{x})$, but why? Consider the goal of minimizing misclassification. Imagine we are free to choose (post inference) the locations of regions R_1 and R_2 which will assign input vectors to either C_1 (in R_1) or C_2 (in R_2). The probability of a mistake integrates over each region, and adds the probability mass of the *other* class(es). This total probability is $P(\text{mistake})$

$$P(\text{mistake}) = P(\mathbf{x} \in R_1, C_1) + P(\mathbf{x} \in R_2, C_1) \tag{1a}$$

$$= \int_{R_1} p(\mathbf{x}, C_2) d\mathbf{x} + \int_{R_2} p(\mathbf{x}, C_1) d\mathbf{x} \tag{1b}$$

I.e. the first integral sums over R_1, but for the model for C_2. If we can choose R_1 and R_2, then if $p(\mathbf{x}, C_1) > p(\mathbf{x}, C_2)$, then this $\mathbf{x}$ should be added to region R_1 (and then $p(\mathbf{x}, C_2$ gets added). This concept is illustrated in Figure 1.

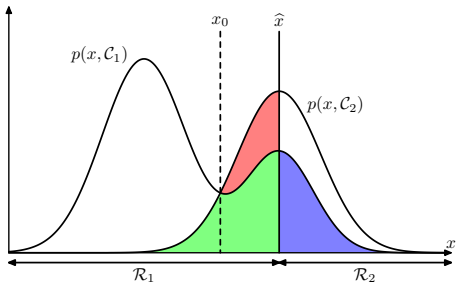

Figure 1. Decision theory for classification (Bishop Figure 1.24). At $\hat{x}$ (a sub-optimal decision threshold), the blue represented the mistakes from class C_1 and the green for class C_2. The red region is the extra error from C_2 being placed in R_1; since $p(x, C_2) > p(x, C_1)$, this region should be in R_2. The optimal decision X_0 is where they cross.

But previously we used the arg max of $P(C_k|\mathbf{x})$, not of the joint, so what is going on? Simply rewriting $P(C_k, \mathbf{x}) = P(C_x|\mathbf{x})p(\mathbf{x})$ shows that they are equivalent because $p(\mathbf{x})$ is a constant for both. We can therefore use probabilistic discriminant models or generative models. This is illustrated in Figure 2.

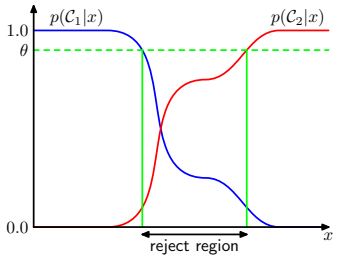

Figure 2. Decision theory for classification (Bishop Figure 1.26). It is equivalent to assign to arg max of $P(C_k, \mathbf{x})$ and $P(C_k|\mathbf{x})$ (plotted here). Note the switch from one region occurs at 1/2. Also plotted (not discussed in text), is a rejection threshold θ that can be used to reject making a decision because we are too uncertain (so perhaps a more complex model or a human must intervene).

3 Minimizing Expected Loss

Consider a few examples of real-world situations where there are costs associated with each mistake (or incorrect action).

Examples:

1. Using an x-ray image to detect cancer. If cancer is diagnosed, but the patient doesn't have cancer, then there is a small loss, say 1, due to the extra tests that will be performed. If cancer is not diagnosed and the patient actually has cancer, then the patient may get sicker and die; this, depending on who you ask, could be a loss of 1000.
2. Evaluating a loan applicant (from R. Neal). We can decide to give the applicant a loan, but if he doesn't pay the loan back, we incur the loss of the loan, say 10. On the other hand, if we decide not to give a loan, and the applicant would actually pay it back, then there is a small loss, say 1, due to lost profits.

It is useful to put the losses into a loss matrix L:

	Cancer (action)	Normal (action)
Cancer	0	1000
Normal	1	0

Table 1. Loss Matrix. Columns are actions. The first row shows $L_{00} = 0$, $L_{01} = 1000$; the second row $L_{10} = 1$, $L_{11} = 0$.

Decision theory states that we should take the **action** that minimizes the **expected loss**:

$$R(a_j|\mathbf{x}) = \sum_k L_{kj} P(C_k|\mathbf{x}) \tag{2}$$

where R is used as for expected loss for taking action a_j ("R" for risk). NB: this is different notation than Bishop (the R and actions part). We therefore need to compute the expected loss for all actions, and then pick action a_j if:

$$\sum_k L_{kj} P(C_k|\mathbf{x}) < \sum_k L_{ki} P(C_k|\mathbf{x}) \qquad \forall i \neq j \tag{3}$$

or more succinctly

$$R(a_j|\mathbf{x}) < R(a_i|\mathbf{x}) \qquad \forall i \neq j \tag{4}$$

For binary classification, the result is simpler: assign to class C_0 if

$$L_{10} P(C_1|\mathbf{x}) < L_{01} P(C_0|\mathbf{x}) \quad \rightarrow \quad \frac{L_{01} P(C_0|\mathbf{x})}{L_{10} P(C_1|\mathbf{x})} > 1 \tag{5}$$

3.1 Example

We run a diagnostic test, revealing $P(C_0|\mathbf{x}) = 0.01$ and $P(C_1|\mathbf{x}) = 0.99$ (i.e. there is a 1% probability of cancer). The actions are a_0: treat as cancer, and a_1: treat as normal. Using the same loss matrix above:

$$R(a_0|\mathbf{x}) = L_{00}P(C_0|\mathbf{x}) + L_{10}P(C_1|\mathbf{x}) \tag{6a}$$
$$= 0 + 1 \cdot 0.99 \approx 1 \tag{6b}$$
$$R(a_1|\mathbf{x}) = L_{01}P(C_0|\mathbf{x}) + L_{11}P(C_1|\mathbf{x}) \tag{6c}$$
$$= 1000 \cdot 0.01 + 0 = 10 \tag{6d}$$

Since $R(a_0|\mathbf{x}) < R(a_1|\mathbf{x})$, we take action a_0, treat like cancer. We can check the binary decision rule:

$$\frac{L_{01}P(C_0|\mathbf{x})}{L_{10}P(C_1|\mathbf{x})} = \frac{1000 \cdot 0.01}{1 \cdot 0.99} = 10 \tag{6e}$$

since $10 > 1$, we take action a_0.