Engineering Applications of Machine Learning and Data Analytics

Decision Making with Bayes and Risk

Contents

1 Overview

Statistical decision theory deals with situations where decisions have to be made under uncertainty, and it aims to provide a rational model to address those scenarios. The Bayesian approach is one of the several ways of formulating such decision problems. 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 with accuracy. It offers a method of formalizing priori beliefs and combining them with the available observations, with the goal of allowing a formal derivation of optimal decision criteria. Although these kinds of situations rarely occur in real life, it gives us an idea to model the optimal classifier against which we can compare other classifiers. Further, it enables us to predict the error we get when we attempt to generalize new patterns. Hence the basic idea goes to choosing the least risky class, i.e. the class for which the expected loss is smallest.

2 Data & Probability Review

2.1 Gaussian Distribution

Throughout the course – and generally – in the practice of machine learning, we will need to be able to demonstrate the effectiveness of our *new groundbreaking approach*. While there are many real-world data sets that we can use, we may sometimes need to show the ability of an algorithm to perform under controlled circumstances. Therefore, we need to know how to generate data from known probability distributions. The Gaussian distribution is – by far – the most popular, which we review in this section. One of the advantages of knowing the exact form of the probability distribution is that we can correctly compute the probabilities. For example, if you proposed in your final project that you have a state-of-the-art method to estimate $p(X|Y)$, where X is a variable and Y is the class label then you'll need to show that your method for obtaining $\widehat{p}(X|Y)$ is close to $p(X|Y)$.

You encountered Gaussian random variables in your previous probability courses; however, in practice we receive more than one variable to make a prediction. Therefore, let us look at the

situation where we have a multivariate Gaussian random variable. A multivariate Gaussian random variable is one where we have p variables that are all Gaussian variables could have covariation. Let $\mathbf{x} \in \mathbb{R}^p$ be distributed as $\mathcal{N}(\mu, \Sigma)$, where $\mathcal{N}(\mu, \Sigma)$ is a Gaussian distribution with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. The probability density for $\mathbf{x}$ is given by

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\mu)\right] \tag{1}$$

where $|\cdot|$ is the determinate. The form of this distribution is not only important for generating data, but for validating algorithms that we will discuss throughout the semester. This is primarily because we typically do not have access to $p(\mathbf{x})$ for a real-world data set with a finite number of resources (e.g., number of samples, computational resources, etc).

2.2 Distances and the Mahalanobis Distance

Many times throughout the semester we will need to either: (a) measure the size of a vector, or (b) measure the distance between two vectors, therefore, let us review some common measures. Let us first consider a vector $\mathbf{x} \in \mathbb{R}^p$ with elements x_i where $i \in [p] := \{1, \ldots, p\}$. The r-norm is formally given by

$$L_r(\mathbf{x}) = \left(\sum_{i=1}^{d} x_i^r\right)^{\frac{1}{r}}$$

which is useful for measuring the size of a vector. This formulation can also be generalized to distances, which the most popular being the *Euclidean Distance*. The Euclidean distance between vectors $\mathbf{x}$ and $\mathbf{z}$ is given by

$$d_2(\mathbf{x}, \mathbf{z}) = \left(\sum_{i=1}^{d} (x_i - z_i)^2\right)^{\frac{1}{2}}$$

The above definitions lend themselves well to measuring the distance between two points; however, what if we want to know the "distance" from a point $\mathbf{x}$ to a probability distribution. It turns out that the Gaussian has this distance hidden in the density function from (1). The *Mahalanobis Distance* is the distance from a point to a distribution. The formal definition of this distance is given by:

$$d_{\text{Mahal}}(\mathbf{x}, \mu, \Sigma) = (\mathbf{x}-\mu)^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\mu)$$

Note that a point $\mathbf{x}$ with "closer" Euclidean distance does not imply a closer Mahalanobis distance with a known μ and Σ.

2.3 Axioms of Probability & Common Properties

At the heart of making predictions, we need a way to cope with uncertainty and probability theory provides us an easy ways to cope with uncertainty. Let us review the Axioms of probability and some rules that will be useful throughout the remainder of this course[1]. A probabilistic model is a mathematical description of an uncertain situation. Think of these uncertain situations as random experiments. They are random in that the outcome is uncertain and is likely to vary upon independent trials of the experiment. The probability triple $(\Omega, \mathcal{F}, P)$:

[1] Read as your lifetime!

- Sample space (Ω): the set of possible outcomes for the random experiment. This set may be finite, countably infinite, or uncountably infinite. The elements of Ω are the individual outcomes ω. The random experiment produces precisely one outcome $\omega \in \Omega$.
- Events ($\mathcal{F}$): the set of possible events for the random experiment. An event is a subset of outcomes. For all discrete models it suffices to consider $\mathcal{F} = P(\Omega)$. For continuous models there is a need for greater mathematical sophistication (measure theory). An event $A \in F$ is a subset of outcomes $A \subset \Omega$.
- Probability (P): a function $P : \mathcal{F} \to [0, 1]$ assigning a number $P(A) \in [0, 1]$ to each $A \in \mathcal{F}$, satisfying the axioms of probability given later in this section. $P(A)$ is the probability of event A being true, i.e., the probability that the random outcome of the experiment, ω, lies in $A : P(\omega \in A)$.

Probability laws The probability $P : \mathcal{F} \to [0, 1]$ in the probability triple $(\Omega, \mathcal{F}, P)$ modeling a random experiment must obey the following three axioms of probability.

- *Non-negativity*: $P(A) \geq 0$ for all $A \in \mathcal{F}$.
- *Additivity*: if A, B are disjoint events ($A \cap B = \emptyset$) then $P(A \cup B) = P(A) + P(B)$. More generally if $|\Omega| = \inf$ and $\{A_1, A_2, \ldots\}$ is an infinite sequence of disjoint events, then $P(A1 \cup A2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$.
- *Normalization*: $P(\Omega) = 1$ (something must happen).

It is also important to realize the properties of independence. Consider the random variables given by A, B and C. If A and B are independent then $P(A, B) = P(A)P(B)$. Furthermore, if random variables A, B are independent given C the $P(A, B|C) = P(A|C)P(B|C)$

Sum Rule The marginal probability of a single random variable can always be obtained by summing (integrating) the probability density function (pdf) over all values of all other variables. For example,

$$P(X) = \sum_{y \in \mathcal{Y}} P(X, Y = y)$$
$$P(X) = \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} P(X, Y = y, Z = z)$$

Product Rule The joint probability can always be obtained by multiplying the conditional probability (conditioned on one of the variables) with the marginal probability of the conditioned variable. For example,

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

Bayes Rule The combination of these two rules gives us Bayes Rule. Let Y denote the outcome that we seek to predict (e.g., healthy or not) and X denote the variable(s) we are provided (e.g., medical records). Then we wish to find

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum_{y \in \mathcal{Y}} P(X, Y = y)} = \frac{P(X|Y)P(Y)}{\sum_{y \in \mathcal{Y}} P(X|Y = y)P(Y = y)}$$

It is the combination of these therms that allows us to develop the strongest decision rule. That is to say, the one that minimize the probability of error or the risk we take when a decision is

made. Each of the terms in this expression has a name associated with it. The prior, $P(Y)$, is our (subjective) degree of belief that the *event* is plausible in the first place. The likelihood, $P(X|Y)$, the likelihood of making an observation, under the condition that the event has occurred. Finally, the evidence, $P(X)$, the probability of making such an observation. It is the combination of these three pieces of information that gives the probability of an event, given that an observations – that provides incomplete information – has been made. The probability computed based on such an observation is then called the posterior probability, $P(Y|X)$.

Example Consider the problem of filtering out spam emails from your inbox. Emails that you receive are either *spam* or *ham*, and they contain a bunch of text. Let $X \in \{0,1\}^W$ which is a vector of size of the number of unique words and a "0" or "1" is placed in the entry of the vector if the work did not or did appear, respectively. This is known as a "bag of words" model (sometimes we'll consider the frequency of a word).

$$P(Y = \text{ham}|Nigerian = 1, Prince = 1, Love = 0, \ldots) = \frac{P(Y = \text{ham})P(Nigerian = 1, Prince = 1, Love = 0, \ldots|y)}{P(Nigerian = 1, Prince = 1, Love = 0, \ldots)}$$

Given our prior knowledge about the words "Nigerian" and "Prince" occurring in spam emails, it will be unlikely that this is a "ham" email.

3 Decision Making with Bayes

The notation for the state of the nature will be denoted by either Y or ω from this point forward. I know if seems confusing, I'll quote the great computer scientist Donald Kunth "*different tasks call for different conventions*[2]".

3.1 Bayes Decision Rule

Bayes theorem can also be used to guide as to which class we should select when we have observed X. The short answer: choose the class ω^* that has the largest posterior probability. More formally, this is given by:

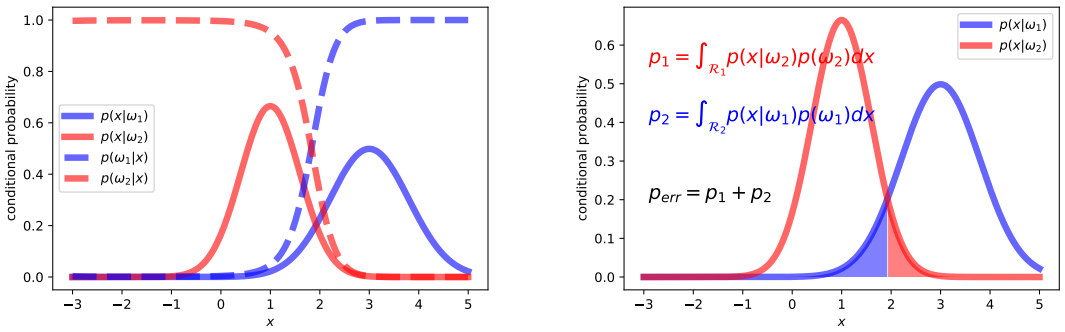
$$\omega^* = \arg\max_{\omega \in \Omega} P(\omega|X) = \arg\max_{\omega \in \Omega} \frac{P(X|\omega)P(\omega)}{P(X)} = \arg\max_{\omega \in \Omega} P(X|\omega)P(\omega)$$

where $P(X)$ can be omitted since it only acts as a normalization constant. Figure 1 shows an example between the likelihood, prior and posterior for two different scenarios.

It can also be shown that the Bayes rule has the minimum probability of error (see later section on risk). Note that this latter point only hold if we can compute all of the terms – likelihood in particular – in Bayes theorem. What if we do not know the parametric form of $P(X|\omega)$? How can we estimate it.

Histogram criteria The simplest estimator of a probability mass function is a histogram. That is to say if we have a two class problem the we can split the data into each class, we can attempt to model $p(\mathbf{x}|Y)$ empirically using the sample that we have. One question: *how much data is enough*? Let us set up a modest histogram criteria to determine when we have enough data to have faith in our estimate of the distribution. For example, let us consider that our advisor told us that we need 30 samples in each bin of the histogram and there are 20 bins.

[2] XKCD reference (`https://xkcd.com/163/`).



(a) Conditional probability (b) Probability of errors

Figure 1: Likelihoods of two different classes, both of which are sampled from Gaussian distributions. Notice that the likelihood is shown with a solid line and a different color for each class. There is also a visualization of the probability of errors.

- That means for 1D data: $20 \times 30 = 600$ instances
- That means for 2D data: $20 \times 20 \times 30 = 12k$ instances
- That means for 3D data: $20 \times 20 \times 30 = 240k$ instances

What!?!?!?! That means with this modest criteria that we would need $240k$ instances to estimate the joint distribution for a 3D variable. Ladies and Gentlemen, meet the *curse of dimensionality*. The fundamental reason for the curse of dimensionality is that modeling high-dimensional data have the potential to be much more complicated than low-dimensional ones, and that those complications are harder to discern. The only way to beat the curse is to incorporate knowledge about the data that is correct (or not).

3.2 The Naïve Bayes Classifier

Estimating the likelihood terms can be extremely burdensome if we do not know the form of the distribution or we simply cannot estimate the term – feasibly – because of the curse of dimensionality. This could be driven by the fact that we simply do not know the dependency relationships among the features. When the dependency relationships among the features used by a classifier are unknown, we generally proceed by taking the simplest assumption, i.e., that the features are conditionally independent given the category[3], i.e.,

$$P(\omega)p(\mathbf{x}|\omega) = P(\omega)\prod_{i=1}^{p} p(x_i|\omega)$$

The maximum likelihood estimator for $P(\omega_j)$.

$$P(\omega_j) = \frac{\#\text{of instances in } \omega_j}{\#\text{of instances in the data set}}$$

where $j \in [c]$.

[3]Note that the notation for p is the number of features (predictors) in a data set and $p(\cdot)$ is a probability mass function.

4 Risk with Bayes

Often we find that not all decisions are associated with equal costs. As an example, while making a decision for loan application, a financial institution takes into account potential losses and gains. In this case, the institution makes sure that an accepted low-risk should increase the profit, while rejected high-risk applicant decreases the loss. In this case, we may find symmetry to be counter-intuitive because loss for a high-risk applicant accepted may be different from the potential gain from a low-risk applicant rejected by mistake. This distorts the concepts of rationality when dealing with decision making using in Bayesian Decision Theory in some domains such as economics and psychology. Consequently, in Bayesian decision making, a well-defined loss or risk function, indicating the potential loss/risk incurred by each possible pair of cause and outcome is critical. Once we decide on loss or risk function, finding the optimal estimate consists of minimizing the expected loss, where the expectation is taken over the posterior distribution over the variable of interest, taking into account any uncertainty over the setting of the variable.

Minimizing Overall Risk To begin this problem, let us consider the possible class outcomes and the actions that we could consider to even make a decision. Before, moving on let us make a few definitions:

- Notations associated with risk:
  - Class, or the state of nature, is denoted by $\omega : \{\omega_1, \omega_2, ..., \omega_c\}$
  - Action α_i is defined as the decision to assign the input to class ω_i. There are ℓ different types of actions: $\{\alpha_1, \alpha_2, ..., \alpha_\ell\}$
  - Cost λ_i is the cost incurred for taking α_i: $\{\lambda_1, \lambda_2, ..., \lambda_\ell\}$

We should know that number of action we would take need not equal the number of classes. What would be one such example when number of actions is not equal to the number of class? One such example can be when we don't any actions; in such case, number of actions is not equal to the number of classes because we can always choose to not to take an action. Our homework forced us to show that the Bayes decision rule gives us the method for minimizing the over all risk (error). Now we look at another perspective to show that this is true.

$$\begin{aligned}\alpha &= \arg\min_{\alpha_i} R(\alpha_i|\mathbf{x}) \\ &= \arg\min_{\alpha_i} \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})\end{aligned}$$

In decision-theoretic terminology, an expected loss is called a risk, and $R(\alpha_i|\mathbf{x})$ 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. We shall now show that this Bayes decision procedure actually provides the optimal performance on an overall risk.

Two Category Classification Example Let us consider these results when applied to the special case of two-category classification problems. Here action α_2 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 given by we obtain

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

where R is the expected risk for taking action α_i given the observed variable $\mathbf{x}$. The fundamental rule is decide ω_1, if

$$R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$$

References

[Alpaydin, 2004] Alpaydin, E. (2004). *Introduction to Machine Learning*. MIT Press.

[Bertsekas and Tsitsiklis, 2008] Bertsekas, D. and Tsitsiklis, J. (2008). *Introduction to Probability*. Athena Scientific.

[Bishop, 2006] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

[Duda et al., 2001] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, Inc., 2nd edition.

[Polikar, 2006] Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45.