Bayes Decision Theory

Alan Yuille

Feb 5 2024

Bayes Introduction

- This lecture introduces Bayes and Bayes Decision Theory
- Bayes Decision Theory
- Empirical Risk
- Critique of Bayes
- Bayes in the Big

Bayes decision theory

- Bayes decision theory (BDT) is a framework for making optimal decisions in the presence of uncertainty. .
- The theory contains three ingredients: (I) A *probability distribution* $P(x, y)$ over the input $x \in \mathcal{X}$ and output $y \in \mathcal{Y}$. (II) A set of *decision rules* $\{\alpha(); \alpha \in \mathcal{A}\}$ where $\alpha(x) \in \mathcal{Y}$. (III) A *loss function* $L(\alpha(x); y)$, which is the cost of making decision $\alpha(x)$ if the real decision should be y.
- The *risk* is specified by the expected loss function $R(\alpha) = \sum_{x,y} P(x, y)L(\alpha(x), y)$.
- The optimal decision is *Bayes rule* $\hat{\alpha} = \arg\min_{\alpha \in \mathcal{A}} R(\alpha)$ minimizing the risk yielding the *Bayes risk* $\min_\alpha R(\alpha) = R(\hat{\alpha})$ (Caveat).

Probabilities: Joint and Conditional Distributions

- By basic probability theory we can re-express the *joint distribution* $P(x, y)$ in two different ways:
- (I) $P(x, y) = P(x|y)P(y)$, where $P(x|y)$ is the *conditional distribution* of x conditioned on y.
- (II) Similarly $P(x, y) = P(y|x)P(x)$.
- By equating $P(x|y)P(y) = P(y|x)P(x)$ we derive *Bayes Theorem* $P(y|x) = P(x|y)P(y)/P(x)$.
- Bayes Theorem is very simple but very important. Particularly for inverse problems like vision where $P(x|y)$ can be interpreted as the probability of generating an observation, e.g., an image, from a world state y. This models the forward process of generating the observation. $P(y|x)$ is the posterior distribution of the world state conditioned on the data.

Probabilities: Likelihoods, Priors, and Posteriors.

- There are three important types of distributions in BDT. The likelihood, the prior (before), and the posterior distribution (after).
- (I) $P(x|y)$ is the *likelihood function* of y and specifies what we know about y given the observation x.
- (II) $P(y)$ specifies the prior knowledge of y independent of the observation.
- (III) $P(y|x)$ is the *posterior distribution* of y after making the observation x, combining the likelihood function and the prior.
- Prior distributions are sometimes controversial (beyond the scope of this course). Here we assume there are ways to estimating them reliably from data.

Bayes Rule

- The expected risk can be re-expressed as $\sum_x P(x) \sum_y P(y|x) L(\alpha(x), y)$.
- Hence the Bayes rule $\hat{\alpha}(.)$ can be expressed as $\hat{\alpha}(x) = \arg\min \sum_y P(y|x) L(\alpha(x), y)$.
- The Bayes rule depends only on the posterior distribution $P(y|x)$. This is an important point that we will return to later.

Bayes Rule and special cases

- Case (I): If the loss function penalizes all errors equally — i.e. $Ł(\alpha(x), y) = \delta(y - \alpha(x))$, where $\delta()$ is the Dirac delta function — then $\hat{\alpha}(x) = \arg\max_y P(y|x)$. This is the *maximum a posteriori* (MAP) estimate of y.
- Case (II). If, in addition, the prior $P(y)$ is the uniform distribution. In this case, $\hat{\alpha}(x) = \arg\max_y P(x|y)$ which is the *maximum likelihood* (ML) estimate of y.

Bayes rule for binary decisions

- The binary case $y \in \{-1, 1\}$ illustrates the trade off between different types of errors. It is conventional to call $y = 1$ the *target* and $y = -1$ as the *distractor*.
- For a decision rule $\alpha(x)$, we define (x, y) to be:
- *false positives* if $\alpha(x) = 1$ and $y = -1$,
- *false negative* if $\alpha(x) = -1$ and $y = 1$.
- A false positive occurs if we predict the input x to be the target when it is a distractor. A false negative occurs if the decision rule predicts the signal to be a distractor but instead it is a target.
- Suppose we are trying to diagnose a disease. We would like a decision rule which is always correct, No false positives or false negatives. In practice this is not possible, and we need to choose a loss function that trades offs the false negatives with the false positives.

Bayes rule for binary decisions: log-likelihood ratio

- For binary decision problems $y \in \{\pm 1\}$, the decision rule can be expressed in terms of the log-likelihood ratio test.
- $\alpha(x) = 1$ if $\log \frac{P(x|y=1)}{P(x|y=-1)} > T$,
- $\alpha(x) = -1$ if $\log \frac{P(x|y=1)}{P(x|y=-1)} < T$,
- The threshold T depends on the prior $p(y)$ and the loss function $L(\alpha(x), y)$ (see next slide).

Bayes rule (III)

- Divide the data (x, y) into four sets:
- (1) the *true positives* $\{(x, y) : \text{ s.t. } \alpha(x) = y = 1\}$;
- (2) the *true negatives* $\{(x, y) : \text{ s.t. } \alpha(x) = y = -1\}$;
- (3) the *false positives* $\{(x, y) : \text{ s.t. } \alpha(x) = 1, y = -1\}$;
- (4) the *false negatives* $\{(x, y) : \text{ s.t. } \alpha(x) = -1, y = 1\}$.
- These four cases correspond to loss function values
  $L(\alpha(x) = 1, y = 1) = T_p$, $L(\alpha(x) = -1, y = -1) = T_n$,
  $L(\alpha(x) = 1, y = -1) = F_p$, $L(\alpha(x) = -1, y = 1) = F_n$.
- Then the best decision rule $\hat{\alpha(.)}$ can be expressed as.

$$\log \frac{P(x|y=1)}{P(x|y=-1)} > \log \frac{T_n - F_p}{T_p - F_n} + \log \frac{P(y=-1)}{P(y=1)}.$$

- The intuition is that the evidence in the log-likelihood must be bigger than our prior biases for the opposite decision, while taking into account the penalties paid for different types of mistakes.

Signal detection theory (I)

A classic book *Signal detection theory and psychophysics* (Green & Swets, 1966) applied BDT to psychophysics. They proposed BDT as a framework for studying perception, particularly auditory.

They typically assumed that the likelihood functions could be expressed as Gaussian distributions.

Consider he two class case, where $y \in \{\pm 1\}$, and suppose that the likelihood functions are specified by Gaussian distributions,
$P(x|y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\{-(x-\mu_y)^2/(2\sigma_y^2)\}$, which differ by their means (μ_1, μ_{-1}) and their variances $(\sigma_1^2, \sigma_{-1}^2)$. The Bayes rule can be expressed in terms of the log-likelihood ratio test:

$$\hat{\alpha}(x) = \arg\max_y y\{-(x-\mu_1)^2/(2\sigma_1^2) - \log\sigma_1 + (x-\mu_{-1})^2/(2\sigma_{-1}^2) + \log\sigma_2 - T\}.$$

Signal detection theory (II)

- This decision rule requires determining whether the data point x is above or below a quadratic polynomial curve in x. In the special case when the standard deviations are identical $\sigma_1^2 = \sigma_2^2$ (so we drop the subscripts $_{1,-1}$), the decision is based only on whether the data point x satisfies:

$$2x(\mu_1 - \mu_{-1}) + (\mu_1^2 - \mu_2^2) < 2T\sigma^2$$

- This special case, with $\sigma_1^2 = \sigma_{-1}^2$, is much studied in signal detection theory (Green & Swets, 1966). It means that the decision is based on a single function $d' = \frac{\mu_1 - \mu_{-1}}{\sigma}$. This quantity is used to quantify human performance for psychophysical tasks.
- Historically Bayes Decision Theory, which was developed during WW2 for like decrypting enemy codes (e.g., the Enigma machine) or for detecting enemy aircraft using radar. Signal Detection theory was arguably the first scientific application of BDT to science. Statisticians resisted BDT, because they disliked the idea of a prior, and it was mostly advocated by mathematicians like I.J. Good who worked with Turing in decrypting codes in WW2 and who, like Turing, wrote papers about AI.

Learning the Probability Distributions

- Bayes Decision Theory assumes that we know the probability distributions $P(x|y)$ and $P(y)$. Or, alternatively the posterior distribution $P(y|x)$.
- If the probability distributions are Gaussian, as in early applications, like Signal Detection Theory, then this reduces to estimating the means and variances of the distributions.
- More generally, BDT can be applied to learn probability distributions from data. We describe this for learning a distribution $P(x)$ of the observed data, but the same approach can be extended to learning conditional distributions $P(x|y)$ and $P(y|x)$.
- We assume functional form for the distribution which is specified by a parameter γ (e.g., a Gaussian distribution where the parameters and the mean and variance). This gives a distribution $P(x|\gamma)$ where γ is treated as random variables to be estimated from the observed data D= $\{\mathrm{x}_i : i = 1, ..., N\}$.

Learning the Probability Distributions

- We assume that the data is *independently identically distributed* (iid). This implied that $P(\mathcal{D}|\gamma) = \prod_{i=1}^{N} P(x_i|\gamma)$.
- The ML estimate $\hat{\gamma}$ is given by $\arg\max_\gamma \prod_{i=1}^{N} P(x_i|\gamma$ or, equivalently, by $\hat{\gamma} = \arg\min_\gamma(-1)\sum_{i=1}^{N} \log P(x_i|\gamma)$.
- This is a special case of BDT because we have assumed a uniform prior $P(\alpha)$ and a loss function where all errors are penalized equally.
- Priors can be used but have limited effect unless the number N of training examples are small. If a prior is used, the MAP estimate is given by $\hat{\gamma} = \arg\min_\gamma(-1)\{\log P(\gamma) + \sum_{i=1}^{N} \log P(x_i|\gamma)$, so the prior has limited effect. There are interesting cognitive science theories for small N (see Griffiths, Chater, Tenenbaum 2024).

The Empirical Risk

- An alternative approach, used in some types of Machine Learning (ML), is to learn the decision rule $\alpha(.)$ directly from the data $\mathcal{D}_N = \{(x_i, y_i) : i = 1, ..., N\}$.
- Advocates argue that data is precious and there is no need to waste it on learning probability distributions if the true goal is to estimate a decision rule.
- This approximates $R(\alpha) = \sum_{x,y} L(\alpha(x), y)P(x, y)$ by $R_{emp,\mathcal{D}_n}(\alpha) = 1/N \sum_{i=1}^{N} L(\alpha(x_i).y_i)$. In the limit as $N \mapsto \infty$ the empirical risk $R_{emp,\mathcal{D}_n} \mapsto R(\alpha)$. This assumes that the $\{(x_i, y_i)\}$ are identically independently distributed (iid) samples from $P(x, y)$. Then we estimate $\hat{\alpha}()$ by minimizing $R_{,\mathcal{D}_n}(\alpha)$.
- A special case is Support Vector Machines (SVMs) which was the most popular ML in computer vision before neural networks. SVM argues that using the data to estimates probabilities is wasteful and it is better to concentrate directly on the decision boundaries. For SVMs this meant using the data to learn the decision boundaries, specified by the *support vectors*.

The Empirical Risk: PAC theory

- There is a beautiful mathematical theory, informally called Probably Approximate Correct (PAC), which gives upper bounds of the amount of data needed for the estimator to be close to $\arg\min R(\alpha)$ depending on the capacity of the decision rule.
- PAC guarantees that, with high probability, the decision rule will *generalize* to new data, provided this comes from the same source (i.e. iid from $P(x, y)$). But these theoretical bounds are not tight and rarely useful in practice. r
- There are two basic concepts underlying PAC. Firstly, *large deviation theory* which quantifies the probability that observations are typical (e.g., if you toss an unbiased coin often enough you will expect a roughly equal number of heads and tails). Secondly, a measure of the *capacity* of the class of decision rules (low capacity means that the decision rules can only represent a limited class of functions).
- PAC theory, and more practical considerations, suggest that you need more training data than the capacity of the set of classifiers $\mathcal{A}$ (regularization can be used to reduce the capacity). But the capacity is very hard to measure, expect for very simple decision rules.

Learning Posterior Probabilities

- Suppose the decision rules can be expressed as $\hat{\alpha}(x) = \arg max_y P(y|x;\alpha)$ where $P(y|x;\alpha)$ is a family of probability distributions parameterized by α. This corresponds to convolutional neural networks (CNNs).
- If we have no prior for α and the loss function penalizes errors equally, the BDT recommends using ML to estimate $\hat{\alpha} = \arg\min -\sum_i \log P(y_i|x_i:\alpha)$, where α is the parameter of a distribution (regression). This corresponds directly to probabilistic learning where we learn the posterior distribution $P(y|x)$ directly, as will be discussed later in the course.
- For some types of decision rules, like some types of deep networks, the capacity is "elastic" and the decision rules generalize well even if there is only a small amount of data and perform increasingly better when there is more. (contrary to what PAC theory suggests). For other classes of decision rules, the capacity can be reduced by regularizing them,
- Almost all loss functions used to train CNNs can be derived from BDT (cross entropy loss is MLE) and sometimes use loss functions (e.g., for edge detection loss functions pay more penalty for missing an edge than for failing to detect an edge).

The Limits of BDP

- BDT measure performance by average case.
- Average case can be problematic if the datasets have biases (as datasets often are) which can be exploited by neural networks. Certain images (e.g., of baby sitting in the road) are underrepresented in the dataset. Context is often biased, e.g., birds are the most likely objects to be in a tree, but many other objects are possible).
- An algorithm trained on data from one source (for counting crowd size in NYC) may perform badly on data from a different source (counting crowd size i Beijing). This is domain shift.
- An alternative strategy is to find the failure modes of algorithms – by searching over the stimulus space – to find the images which cause the algorithms to fail. Adversarial examiners, which will be discussed later in the course.
- If BDT is sufficient – i.e., we can construct an enormous dataset which can represent all possible images (for training and testing algorithms) – then regression methods that estimate $P(y|x)$ are sufficient. But if not, and if we have an *open world* where we keep seeing new images that differ from those in our training sets, then we will Bayes $P(x|y)$ and $P(y)$.