Machine Learning in Bioinformatics

Introduction to Decision Theory

Philipp Benner
philipp.benner@bam.de

VP.1 - eScience
Federal Institute of Materials Research and Testing (BAM)

April 25, 2024

Decision Theory: Outline

- Decision theory: choice under uncertainty

Decision Theory: Outline

- Decision theory: choice under uncertainty
- Hypothesis testing and parameter estimation are special cases of decision theory [Wald, 1939]:

Decision Theory: Outline

- Decision theory: choice under uncertainty
- Hypothesis testing and parameter estimation are special cases of decision theory [Wald, 1939]:
- Three components of decision theory
  - Assignment of probabilities to events
    - Bayes theorem
    - Maximum entropy approach
  - A loss function that describes the cost of a decision
  - A rule for selecting the best decision

- Probability theory is not just a set of rules for computing frequencies

- Probability theory is not just a set of rules for computing frequencies
- Probability theory is the calculus of inductive reasoning as proposed by Laplace [Good, 1950, Savage, 1972]

Decision Theory: Historical Context / Literature

- Probability theory is not just a set of rules for computing frequencies
- Probability theory is the calculus of inductive reasoning as proposed by Laplace [Good, 1950, Savage, 1972]
- It is seen as an extension of logic calculus [Jaynes, 2003]

- Probability theory is not just a set of rules for computing frequencies
- Probability theory is the calculus of inductive reasoning as proposed by Laplace [Good, 1950, Savage, 1972]
- It is seen as an extension of logic calculus [Jaynes, 2003]
- Allows the assignment of (subjective) probabilities to propositions or events (i.e. *the states of nature*) to quantify their plausibility

- Probability theory is not just a set of rules for computing frequencies
- Probability theory is the calculus of inductive reasoning as proposed by Laplace [Good, 1950, Savage, 1972]
- It is seen as an extension of logic calculus [Jaynes, 2003]
- Allows the assignment of (subjective) probabilities to propositions or events (i.e. *the states of nature*) to quantify their plausibility
- For instance, the plausibility of proposition A given some other proposition B is true $(A \mid B)$

The Running Example

Widget factory [Jaynes, 1963]

Mr. A is in charge of a Widget factory. Every morning he must decide whether to paint the daily run of 200 widgets red, yellow, or green. He does not know how many orders for each type will come in during the day. However, the promise of the factory is that it can make delivery on any size order within 24 hours. This is of course not realistic, but Mr. A's job is to fulfill this promise as best as he can.

Running Example

A priori knowledge:

- Mr. A arrives at work an notices that there are 100 red, 150 yellow, and 50 green widgets in stock

Running Example

A priori knowledge:

- Mr. A arrives at work an notices that there are 100 red, 150 yellow, and 50 green widgets in stock
- In addition, Mr. A learns that the expected daily orders of widgets are 50 for red, 100 for yellow and 10 for green

Running Example

Every morning Mr. A has to choose among three possible decisions:

- D_1 = "make red widgets today"
- D_2 = "make yellow widgets today"
- D_3 = "make green widgets today"

Running Example

Every morning Mr. A has to choose among three possible decisions:

- D_1 = "make red widgets today"
- D_2 = "make yellow widgets today"
- D_3 = "make green widgets today"

We discuss two *decision problems*:

1. Which decision is optimal?
2. How to estimate the expected daily orders?

Problem 1: Optimal Decision

Probability Distribution: Random Variables

We define three random variables X_1, X_2, and X_3 for the total daily ordered number of red, yellow, and green widgets, respectively.

We define three random variables X_1, X_2, and X_3 for the total daily ordered number of red, yellow, and green widgets, respectively.

- A probability distribution over X_1, X_2 and X_3

We define three random variables X_1, X_2, and X_3 for the total daily ordered number of red, yellow, and green widgets, respectively.

- A probability distribution over X_1, X_2 and X_3
- Our prior knowledge is

$$\mathbb{E}\, X_1 = 100, \mathbb{E}\, X_2 = 150, \text{and}\, \mathbb{E}\, X_3 = 50$$

We define three random variables X_1, X_2, and X_3 for the total daily ordered number of red, yellow, and green widgets, respectively.

- A probability distribution over X_1, X_2 and X_3
- Our prior knowledge is

$$\mathbb{E}\, X_1 = 100, \mathbb{E}\, X_2 = 150, \text{and}\, \mathbb{E}\, X_3 = 50$$

- How do we get from our prior knowledge to a probability distribution?

Probability Distribution: Entropy

Entropy measures the uncertainty about the outcomes of a random variable X.

Entropy measures the uncertainty about the outcomes of a random variable X.

Entropy [Shannon, 1948]

Let X be a discrete random variable, then the entropy H of X is given by

$$H(X) = -\sum_x \mathrm{pr}(X = x) \log \mathrm{pr}(X = x)\,.$$

Entropy measures the uncertainty about the outcomes of a random variable X.

Entropy [Shannon, 1948]

Let X be a discrete random variable, then the entropy H of X is given by

$$H(X) = -\sum_{x} \mathrm{pr}(X = x) \log \mathrm{pr}(X = x)\,.$$

If X is a continuous random variable with density f_X, then

$$H(X) = -\int_{x} f_X(x) \log f_X(x) \mathrm{d}x\,.$$

For densities $H(X)$ can be negative!

Probability Distribution: Entropy of Bernoulli Trials

Consider a single coin flip, which we also call a Bernoulli trial. In this case, X is discrete and can take two values, either head (x_1) or tail (x_2). Furthermore, we define $\mathrm{pr}(X = x_1) = p$ so that $\mathrm{pr}(X = x_2) = 1 - p$.

Probability Distribution: Entropy of Bernoulli Trials

Consider a single coin flip, which we also call a Bernoulli trial. In this case, X is discrete and can take two values, either head (x_1) or tail (x_2). Furthermore, we define $\mathrm{pr}(X = x_1) = p$ so that $\mathrm{pr}(X = x_2) = 1 - p$.

The entropy $H(X)$ given by

$$H(X) = -p \log p - (1 - p) \log(1 - p) .$$

Probability Distribution: Entropy of Bernoulli Trials

Consider a single coin flip, which we also call a Bernoulli trial. In this case, X is discrete and can take two values, either head (x_1) or tail (x_2). Furthermore, we define $\mathrm{pr}(X = x_1) = p$ so that $\mathrm{pr}(X = x_2) = 1 - p$.

The entropy $H(X)$ given by

$$H(X) = -p \log p - (1 - p) \log(1 - p) \,.$$

Examples:

- The entropy is maximal (i.e. $H(X) \approx 0.693$) if $p = 0.5$, becasue we are most uncertain about the outcome of the coin flip.

Probability Distribution: Entropy of Bernoulli Trials

Consider a single coin flip, which we also call a Bernoulli trial. In this case, X is discrete and can take two values, either head (x_1) or tail (x_2). Furthermore, we define $\mathrm{pr}(X = x_1) = p$ so that $\mathrm{pr}(X = x_2) = 1 - p$.

The entropy $H(X)$ given by

$$H(X) = -p \log p - (1 - p) \log(1 - p) \,.$$

Examples:

- The entropy is maximal (i.e. $H(X) \approx 0.693$) if $p = 0.5$, becasue we are most uncertain about the outcome of the coin flip.
- The entropy is minimal (i.e. $H(X) = 0$) if $p = 0$ or $p = 1$, because we can be sure about the outcome of the coin flip.

Probability Distribution: Maximum Entropy

The maximum entroy approach is a method to select a probability distribution such that it reflects our prior knowledge, without assuming anything beyond that.

The maximum entroy approach is a method to select a probability distribution such that it reflects our prior knowledge, without assuming anything beyond that.

Maximum Entropy Approach

Let $T(X) = (T_1(X), \ldots, T_m(X)) \in \mathbb{R}^m$ denote a statistic of X. Assume that $\theta = \mathbb{E}\, T(X)$ is given. The maximum entropy approach determines the distribution of X as the maximizer of the following optimization program:

$$\begin{aligned} \text{maximize} \quad & H(X) \\ \text{subject to} \quad & \mathbb{E}\, T(X) = \theta \end{aligned}$$

We regard the constraints $T(X) = \theta$ as our prior knowledge.

The maximum entroy approach is a method to select a probability distribution such that it reflects our prior knowledge, without assuming anything beyond that.

Maximum Entropy Approach

Let $T(X) = (T_1(X), \dots, T_m(X)) \in \mathbb{R}^m$ denote a statistic of X. Assume that $\theta = \mathbb{E}\, T(X)$ is given. The maximum entropy approach determines the distribution of X as the maximizer of the following optimization program:

$$\begin{aligned} \text{maximize} \quad & H(X) \\ \text{subject to} \quad & \mathbb{E}\, T(X) = \theta \end{aligned}$$

We regard the constraints $T(X) = \theta$ as our prior knowledge.

In modern terms we also call $T(X)$ the features of our observed data.

- Maximum entropy distribution over a discrete set $\{0, 1, \ldots n\}$ with no constraints:

$$\mathrm{pr}(x) = \frac{1}{n+1} \qquad \text{(uniform)}$$

- Maximum entropy distribution over a discrete set $\{0, 1, \ldots n\}$ with no constraints:

$$\mathrm{pr}(x) = \frac{1}{n+1} \qquad \text{(uniform)}$$

- Maximum entropy distribution on $\mathbb{R}^+$ with known mean μ:

$$\mathrm{pr}(x) = \frac{1}{\mu} \exp\left\{-\frac{X}{\mu}\right\} \qquad \text{(exponential)}$$

Probability Distribution: MaxEnt Distributions

- Maximum entropy distribution over a discrete set $\{0, 1, \ldots n\}$ with no constraints:

$$\mathrm{pr}(x) = \frac{1}{n+1} \qquad \text{(uniform)}$$

- Maximum entropy distribution on $\mathbb{R}^+$ with known mean μ:

$$\mathrm{pr}(x) = \frac{1}{\mu} \exp\left\{-\frac{X}{\mu}\right\} \qquad \text{(exponential)}$$

- Maximum entropy distribution on $\mathbb{R}$ with known mean μ and variance σ^2:

$$\mathrm{pr}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \qquad \text{(normal)}$$

Probability Distribution: MaxEnt Distributions

- Maximum entropy distribution over a discrete set $\{0, 1, \ldots n\}$ with no constraints:

$$\mathrm{pr}(x) = \frac{1}{n+1} \qquad \text{(uniform)}$$

- Maximum entropy distribution on $\mathbb{R}^+$ with known mean μ:

$$\mathrm{pr}(x) = \frac{1}{\mu} \exp\left\{-\frac{X}{\mu}\right\} \qquad \text{(exponential)}$$

- Maximum entropy distribution on $\mathbb{R}$ with known mean μ and variance σ^2:

$$\mathrm{pr}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \qquad \text{(normal)}$$

- In general, maximum entropy distributions are members of the exponential family

We know the expected daily orders θ_j for all colors j.

$$\begin{aligned} \text{maximize} \quad & H(X_j) \\ \text{subject to} \quad & \mathbb{E}\, X_j = \theta_j \end{aligned}$$

Probability Distribution: Running Example

We know the expected daily orders θ_j for all colors j.

$$\begin{aligned} \text{maximize} \quad & H(X_j) \\ \text{subject to} \quad & \mathbb{E}\, X_j = \theta_j \end{aligned}$$

The solution to this optimization problem is given by

$$\mathrm{pr}(X_j = k) = \frac{1}{1+\theta_j}\left(\frac{\theta_j}{\theta_j + 1}\right)^k \qquad \text{(geometric distribution)}$$

Probability Distribution: Running Example

We know the expected daily orders θ_j for all colors j.

$$\begin{aligned} \text{maximize} \quad & H(X_j) \\ \text{subject to} \quad & \mathbb{E} X_j = \theta_j \end{aligned}$$

The solution to this optimization problem is given by

$$\mathrm{pr}(X_j = k) = \frac{1}{1+\theta_j} \left(\frac{\theta_j}{\theta_j + 1} \right)^k \qquad \text{(geometric distribution)}$$

(see backup slides!). More specifically, we have

$$\begin{aligned} \mathrm{pr}(X_1 = k) &= \tfrac{1}{101} \left(\tfrac{100}{101} \right)^k \quad \text{(red)} \\ \mathrm{pr}(X_2 = k) &= \tfrac{1}{151} \left(\tfrac{150}{151} \right)^k \quad \text{(yellow)} \\ \mathrm{pr}(X_3 = k) &= \tfrac{1}{51} \left(\tfrac{50}{51} \right)^k \quad \text{(green)} \end{aligned}$$

Our goal is to make a decision under uncertain future states of nature

Loss Function: Motivation

Our goal is to make a decision under uncertain future states of nature

A probability distribution allows to assign probabilities to events or states of nature.

Loss Function: Motivation

Our goal is to make a decision under uncertain future states of nature

A probability distribution allows to assign probabilities to events or states of nature.

For instance, Mr. A may evaluate the (subjective) probability of receiving x_1 daily orders for red, x_2 for yellow, and x_3 for green widgets.

Loss Function: Motivation

Our goal is to make a decision under uncertain future states of nature

A probability distribution allows to assign probabilities to events or states of nature.

For instance, Mr. A may evaluate the (subjective) probability of receiving x_1 daily orders for red, x_2 for yellow, and x_3 for green widgets.

A probability distribution alone does not allow Mr. A to decide what widgets to produce.

Loss Function

A loss function L quantifies the loss when making decision D and x turns out to be the true state of nature.

Loss Function

A loss function L quantifies the loss when making decision D and x turns out to be the true state of nature.

For instance, for Mr. A the loss might be equal to the number of widgets he cannot deliver. Remember that he has stock of $S_1 = 100$ red, $S_2 = 150$, and $S_3 = 50$ widgets.

Loss Function

A loss function L quantifies the loss when making decision D and x turns out to be the true state of nature.

For instance, for Mr. A the loss might be equal to the number of widgets he cannot deliver. Remember that he has stock of $S_1 = 100$ red, $S_2 = 150$, and $S_3 = 50$ widgets.

If he decides to produce red widgets today (D_1) and x_1, x_2, x_3 represent the daily order of red, yellow, and green widgets, then the loss function is

$$L(D_1; x_1, x_2, x_3) = (x_1 - S_1 - 200)^+ + (x_2 - S_2)^+ + (x_3 - S_3)^+$$

where $(x)^+ = \max(0, x)$.

- Component 1: Probability distribution

$$\mathrm{pr}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = p_1(x_1)p_2(x_2)p(x_3)$$

Decision Rule: Combining Probability and Loss

- Component 1: Probability distribution

$$\mathrm{pr}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = p_1(x_1)p_2(x_2)p(x_3)$$

- Component 2: Loss function

$$L(D_j; x_1, x_2, x_3)$$

Decision Rule: Combining Probability and Loss

- Component 1: Probability distribution

$$\mathrm{pr}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = p_1(x_1)p_2(x_2)p(x_3)$$

- Component 2: Loss function

$$L(D_j; x_1, x_2, x_3)$$

- Component 1 and 2: Weighted loss

$$L(D_j; x_1, x_2, x_3)p_1(x_1)p_2(x_2)p(x_3)$$

Decision Rule: Combining Probability and Loss

- Component 1: Probability distribution

$$\mathrm{pr}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = p_1(x_1)p_2(x_2)p(x_3)$$

- Component 2: Loss function

$$L(D_j; x_1, x_2, x_3)$$

- Component 1 and 2: Weighted loss

$$L(D_j; x_1, x_2, x_3)p_1(x_1)p_2(x_2)p(x_3)$$

- We do not know x_1, x_2, and x_3! Expected loss

$$L(D_j) = \sum_{x_1, x_2, x_3} L(D_j; x_1, x_2, x_3)p_1(x_1)p_2(x_2)p(x_3)$$

Decision Rule: Combining Probability and Loss

- Component 1: Probability distribution

$$\mathrm{pr}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = p_1(x_1)p_2(x_2)p(x_3)$$

- Component 2: Loss function

$$L(D_j; x_1, x_2, x_3)$$

- Component 1 and 2: Weighted loss

$$L(D_j; x_1, x_2, x_3)p_1(x_1)p_2(x_2)p(x_3)$$

- We do not know x_1, x_2, and x_3! Expected loss

$$L(D_j) = \sum_{x_1, x_2, x_3} L(D_j; x_1, x_2, x_3)p_1(x_1)p_2(x_2)p(x_3)$$

- Decision rule: Minimum expected loss $\hat{D} = \min_{D_j} L(D_j)$

Decision Rule: Combinding Probability and Loss

Mr. A computed the following expected losses:

- $L(D_1) = 22.4$ $(widgets)$ for producing red widgets
- $L(D_2) = 9.7$ $(widgets)$ for producing yellow widgets
- $L(D_3) = 28.9$ $(widgets)$ for producing greed widgets

To minimize the expected loss, Mr. A decides to produce yellow widgets today!

Problem 2:
Estimation of Daily Orders

Probability Distribution: Bayes Theorem

Assume we know the daily orders

$$x_{1j},\ x_{2j},\ \ldots,\ x_{nj}$$

for the past n days for color j.

Probability Distribution: Bayes Theorem

Assume we know the daily orders

$$x_{1j},\ x_{2j},\ \ldots,\ x_{nj}$$

for the past n days for color j.

We introduce a random variable for X_{ij} for the total orders at day i for widgets of color j.

Probability Distribution: Bayes Theorem

Assume we know the daily orders

$$x_{1j},\ x_{2j},\ \ldots,\ x_{nj}$$

for the past n days for color j.

We introduce a random variable for X_{ij} for the total orders at day i for widgets of color j.

For simplicity we write $\{\bar{X}_j = \bar{x}_j\} = \{X_{ij} = x_{ij}\}_j$

Probability Distribution: Bayes Theorem

Let θ_j denote the expected daily number of orders for color j, then by assuming independence and from our derivation of *Problem 1* the likelihood is given by

$$\begin{aligned}
\mathrm{pr}(\bar{X}_j = \bar{x}_j \mid \Theta_j = \theta_j) &= \prod_{i=1}^{n} \frac{1}{1+\theta_j} \left(\frac{\theta_j}{\theta_j + 1} \right)^{x_{ij}} \\
&= \prod_{i=1}^{n} \varphi_j (1 - \varphi_j)^{x_{ij}}
\end{aligned}$$

where $\varphi_j = \frac{1}{1+\theta_j}$.

Probability Distribution: Bayes Theorem

Let θ_j denote the expected daily number of orders for color j, then by assuming independence and from our derivation of *Problem 1* the likelihood is given by

$$\mathrm{pr}(\bar{X}_j = \bar{x}_j \mid \Theta_j = \theta_j) = \prod_{i=1}^{n} \frac{1}{1+\theta_j} \left(\frac{\theta_j}{\theta_j + 1} \right)^{x_{ij}}$$
$$= \prod_{i=1}^{n} \varphi_j (1 - \varphi_j)^{x_{ij}}$$

where $\varphi_j = \frac{1}{1+\theta_j}$.

However, in order to decide for a particular θ, we need

$$\mathrm{pr}(\Theta = \theta \mid \bar{X}_j = \bar{x}_j) = \quad ?$$

Bayes Theorem (inverse probability) [Bayes and Price, 1763, Laplace, 1774]

Let X and Θ denote two random variables, where X typically represents the observed data and Θ the parameter or hypothesis of interest. Bayes theorem is given by

$$\mathrm{pr}(\Theta = \theta \mid X = x) = \frac{\mathrm{pr}(X = x, \Theta = \theta)}{\mathrm{pr}(X = x)} = \frac{\mathrm{pr}(X = x \mid \Theta = \theta)\mathrm{pr}(\Theta = \theta)}{\mathrm{pr}(X = x)}$$

where

$\mathrm{pr}(\Theta = \theta \mid X = x)$	is called the posterior distribution,
$\mathrm{pr}(X = x \mid \Theta = \theta)$	the likelihood,
$\mathrm{pr}(\Theta = \theta)$	the prior distribution, and
$\mathrm{pr}(X = x)$	the marginal likelihood or evidence.

Probability Distribution: Bayes Theorem

As prior we select the maximum entropy distribution (Beta distribution)

$$\mathrm{pr}(\Phi = \varphi) = \frac{1}{\mathrm{Beta}(\alpha\ , \beta\ )} \varphi^{\alpha\ -1}(1-\varphi)^{\beta\ -1}$$

for pseudocounts α and β.

$$\Rightarrow \quad \mathrm{pr}(\Phi = \varphi \,|\, \bar{X}_j = \bar{x}_j) = \frac{1}{\mathrm{Beta}(\alpha', \beta')} \varphi^{\alpha'-1}(1-\varphi)^{\beta'-1}$$

where $\alpha' = \alpha + n$, and $\beta' = \beta + \sum_i x_{ij}$.

As prior we select the maximum entropy distribution (Beta distribution)

$$\mathrm{pr}(\Phi = \varphi) = \frac{1}{\mathrm{Beta}(\alpha\ , \beta\ )} \varphi^{\alpha\ -1}(1 - \varphi)^{\beta\ -1}$$

for pseudocounts α and β.

$$\Rightarrow \quad \mathrm{pr}(\Phi = \varphi \,|\, \bar{X}_j = \bar{x}_j) = \frac{1}{\mathrm{Beta}(\alpha', \beta')} \varphi^{\alpha'-1}(1 - \varphi)^{\beta'-1}$$

where $\alpha' = \alpha + n$, and $\beta' = \beta + \sum_i x_{ij}$.

The prior is called conjugate when the posterior is of the same form

Probability Distribution: Remarks

- For complex applications, there often is no conjugate prior

- For complex applications, there often is no conjugate prior
- In such cases, the posterior might not have an analytical solution

Probability Distribution: Remarks

- For complex applications, there often is no conjugate prior
- In such cases, the posterior might not have an analytical solution
- There exist several methods to approximate the posterior distribution
  - Laplace approximation
  - Variational Bayes
  - Metropolis Hastings (MC) and Markov chain Metropolis Hastings (MCMC) methods
  - ...

Loss Function: Running Example

A common loss function for parameter estimation is the squared error loss

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$

where $\hat{\theta}$ denotes our estimate and θ the true parameter.

Loss Function: Running Example

A common loss function for parameter estimation is the squared error loss

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$

where $\hat{\theta}$ denotes our estimate and θ the true parameter.

A computationally attractive choice is the 0-1-loss

$$L_{01}(\hat{\theta}, \theta) = \begin{cases} 0 & \text{if } \hat{\theta} \neq \theta \\ 1 & \text{if } \hat{\theta} = \theta \end{cases}$$

which leads to the maximum a posteriori estimate when using the minimum expected loss as decision rule

Solving the minimum expected loss for different loss functions:

- Squared error loss

$$\begin{aligned}\hat{\theta} &= \arg\min_{\hat{\theta}} \int (\hat{\theta} - \theta)^2 \mathrm{pr}(\Theta = \theta \,|\, X = x) \mathrm{d}\theta \\ &= \int \theta \, \mathrm{pr}(\Theta = \theta \,|\, X = x) \mathrm{d}\theta \\ &= \mathbb{E}\left[\Theta \,|\, X = x\right]\end{aligned}$$

Solving the minimum expected loss for different loss functions:

- Squared error loss

$$\begin{aligned}\hat{\theta} &= \arg\min_{\hat{\theta}} \int (\hat{\theta} - \theta)^2 \mathrm{pr}(\Theta = \theta \,|\, X = x) \mathrm{d}\theta \\ &= \int \theta \, \mathrm{pr}(\Theta = \theta \,|\, X = x) \mathrm{d}\theta \\ &= \mathbb{E}\left[\Theta \,|\, X = x\right]\end{aligned}$$

- 0-1-loss

$$\begin{aligned}\hat{\theta} &= \arg\min_{\hat{\theta}} \int L_{01}(\hat{\theta}, \theta) \mathrm{pr}(\Theta = \theta \,|\, X = x) \mathrm{d}\theta \\ &= \arg\max_{\theta} \quad \mathrm{pr}(\Theta = \theta \,|\, X = x) \qquad \text{(MAP)}\end{aligned}$$

Decision Rule: Minimax Principle

- The minimax decision rule is used to minimize the possible loss for worst case scenarios

Decision Rule: Minimax Principle

- The minimax decision rule is used to minimize the possible loss for worst case scenarios
- In game theory, the minimax decision rule is used because we know the opponent will try to maximize our loss

Decision Rule: Minimax Principle

- The minimax decision rule is used to minimize the possible loss for worst case scenarios
- In game theory, the minimax decision rule is used because we know the opponent will try to maximize our loss
- If Mr. A applied the minimax principle, he would choose the following decision

$$\hat{D} = \arg\min_{D_j} \max_{x_1, x_2, x_3} L(D_j; x_1, x_2, x_3)$$

Decision Rule: Minimax Principle

- The minimax decision rule is used to minimize the possible loss for worst case scenarios
- In game theory, the minimax decision rule is used because we know the opponent will try to maximize our loss
- If Mr. A applied the minimax principle, he would choose the following decision

$$\hat{D} = \arg\min_{D_j} \max_{x_1, x_2, x_3} L(D_j; x_1, x_2, x_3)$$

- For Mr. A, the minimax principle is unrealistic, because nature is not actively playing against him

References I

- Bayes, T. and Price, R. (1763).
  An essay towards solving a problem in the doctrine of chances.
  Philosophical Transactions of the Royal Society of London, 53:370–418.
- Boyd, S. and Vandenberghe, L. (2004).
  Convex optimization.
  Cambridge university press.
- Good, I. J. (1950).
  Probability and the weighing of evidence.
  Technical report, C. Griffin London.
- Jaynes, E. (1963).
  New engineering applications of information theory.
  Available at `https://bayes.wustl.edu/etj/node1.html`.
- Jaynes, E. T. (2003).
  Probability theory: The logic of science.
  Cambridge university press.

References II

Laplace, P.-S. (1774).
Mémoire sur la probabilité des causes par les événements.
Mém. de math. et phys. présentés à l'Acad. roy. des sci, 6:621–656.

Savage, L. J. (1972).
The foundations of statistics.
Courier Corporation.

Shannon, C. E. (1948).
A mathematical theory of communication.
The Bell system technical journal, 27(3):379–423.

Wald, A. (1939).
Contributions to the theory of statistical estimation and testing hypotheses.
The Annals of Mathematical Statistics, 10(4):299–326.

Derivation of MaxEnt Distribution

Maximum Entropy Distribution

The distribution of X is determined by the following optimization program

$$\begin{aligned} \text{maximize} \quad & H(X) \\ \text{subject to} \quad & \mathbb{E}X = \theta \end{aligned}$$

The Lagrangian is

$$L(p, \lambda_0, \lambda_1) = -\sum_{k=1}^{\infty} p_k \log p_k - \lambda_0 \left(\sum_{k=0}^{\infty} p_k - 1 \right) - \lambda_1 \left(\sum_{k=0}^{\infty} k p_k - \theta \right)$$

By differentiating with respect to p_k we obtain

$$\begin{aligned} p_k &= \exp\{-\lambda_0 - \lambda_1 k - 1\} \\ &\equiv \exp\{-\lambda_0 - \lambda_1 k\} \end{aligned}$$

Maximum Entropy Distribution

The Lagrangian multipliers are determined by the constraints. For λ_0 we have

$$\sum_{k=1}^{\infty} p_k = 1 \Rightarrow \sum_{k=1}^{\infty} \exp\left\{-\lambda_0 - \lambda_1 k\right\} = 1$$

From which it follows that

$$\lambda_0 = -\log\left(1 - \exp\left(-\lambda_1\right)\right)$$

Furthermore, for λ_1 we have

$$\sum_{k=1}^{\infty} k p_k = \theta \Rightarrow \sum_{k=1}^{\infty} k \exp\left\{-\lambda_0 - \lambda_1 k\right\} = \theta$$

so that

$$(1 - e^{-\lambda_1}) \frac{e^{\lambda_1}}{(e_1^{\lambda} - 1)^2} = \frac{1}{e_1^{\lambda} - 1} = \theta$$

Maximum Entropy Distribution

It follows that

$$\lambda_1 = \log\left(\frac{1}{\theta} + 1\right)$$

As a result, we have

$$\begin{aligned} p_k &= \exp\{-\lambda_0 - \lambda_1 k\} \\ &= \frac{1}{(1+\theta)(1+\frac{1}{\theta})^k} \\ &= \frac{1}{1+\theta}\left(\frac{\theta}{\theta+1}\right)^k \end{aligned}$$

MaxEnt Distribution with Additional Prior Knowledge

Maximum Entropy Distribution

Additional prior knowledge:

- Mr. A learns that the avarage individual order is 75 for red, 10 for yellow, and 20 for green widgets

Maximum Entropy Distribution

We know the expected number of daily orders θ_j and the expected number of individual orders ϕ_j for all colors j.

Maximum Entropy Distribution

We know the expected number of daily orders θ_j and the expected number of individual orders ϕ_j for all colors j.

In addition to X_j we define a new random variable Y_j, i.e.

$$
\begin{aligned}
X_j &= n && : \text{daily order of size } n \text{ for color } j \\
Y_j &= m && : m \text{ individual orders for color } j \text{ per day}
\end{aligned}
$$

so that $\mathbb{E}\, X_j = \theta_j$ and $\mathbb{E}\, Y_j = \phi_j$.

Maximum Entropy Distribution

We know the expected number of daily orders θ_j and the expected number of individual orders ϕ_j for all colors j.

In addition to X_j we define a new random variable Y_j, i.e.

$$\begin{aligned} X_j = n \quad &: \text{daily order of size } n \text{ for color } j \\ Y_j = m \quad &: m \text{ individual orders for color } j \text{ per day} \end{aligned}$$

so that $\mathbb{E}\, X_j = \theta_j$ and $\mathbb{E}\, Y_j = \phi_j$.

Furthermore, we link X_j and Y_j through a third random variable Z_{ij}, which denotes the number of individual orders of size j for color j.

$$X_j = \sum_{j=1}^{\infty} j Z_{ij}\,, \quad Y_j = \sum_{j=1}^{\infty} Z_{ij}$$

Maximum Entropy Distribution

The maximum entropy problem is given by

$$\begin{aligned} \text{maximize} \quad & H(Z_{i1}, Z_{i2}, \dots) \\ \text{subject to} \quad & \mathbb{E}\, X_j = \theta_j \\ & \mathbb{E}\, Y_j = \phi_j \end{aligned}$$

Maximum Entropy Distribution

The maximum entropy problem is given by

$$\begin{aligned} \text{maximize} \quad & H(Z_{i1}, Z_{i2}, \dots) \\ \text{subject to} \quad & \mathbb{E}\, X_j = \theta_j \\ & \mathbb{E}\, Y_j = \phi_j \end{aligned}$$

What is the solution to this optimization problem?

Maximum Entropy Distribution

The maximum entropy problem is given by

$$\begin{aligned} \text{maximize} \quad & H(Z_{i1}, Z_{i2}, \ldots) \\ \text{subject to} \quad & \mathbb{E}\, X_j = \theta_j \\ & \mathbb{E}\, Y_j = \phi_j \end{aligned}$$

What is the solution to this optimization problem?

Analytically already hard to solve.