A brief overview of probability theory in data science

Geert Verdoolaege
1Department of Applied Physics, Ghent University, Ghent, Belgium
2Laboratory for Plasma Physics, Royal Military Academy (LPP–ERM/KMS), Brussels, Belgium

Tutorial 3rd IAEA Technical Meeting on Fusion Data Processing, Validation and Analysis, 27-05-2019
Overview

1. Origins of probability
2. Frequentist methods and statistics
3. Principles of Bayesian probability theory
4. Monte Carlo computational methods
5. Applications
   - Classification
   - Regression analysis
6. Conclusions and references
Overview

1 Origins of probability
2 Frequentist methods and statistics
3 Principles of Bayesian probability theory
4 Monte Carlo computational methods
5 Applications
   • Classification
   • Regression analysis
6 Conclusions and references
Early history of probability

- Earliest traces in Western civilization: Jewish writings, Aristotle

- Notion of probability in law, based on evidence

- Usage in finance

- Usage and demonstration in gambling
World is knowable but uncertainty due to human ignorance

William of Ockham: *Ockham's razor*

*Probabilis*: a supposedly ‘provable’ opinion

Counting of authorities

Later: degree of truth, a scale

Quantification:

- Law, faith → Bayesian notion
- Gaming → frequentist notion
Quantification

- 17th century: Pascal, Fermat, Huygens
- Comparative testing of hypotheses
- Population statistics
- 1713: Ars Conjectandi by Jacob Bernoulli:
  - Weak law of large numbers
  - Principle of indifference
- De Moivre (1718): The Doctrine of Chances
Bayes and Laplace

- Paper by Thomas Bayes (1763): inversion of binomial distribution
- Pierre Simon Laplace:
  - Practical applications in physical sciences
  - 1812: *Théorie Analytique des Probabilités*
Frequentism and sampling theory

- George Boole (1854), John Venn (1866)
- Sampling from a ‘population’
- Notion of ‘ensembles’
- Andrey Kolmogorov (1930s): axioms of probability
- Applications in social sciences, medicine, natural sciences
Maximum entropy and Bayesian methods

- Statistical mechanics with Ludwig Boltzmann (1868): maximum entropy energy distribution
- Josiah Willard Gibbs (1874–1878)
- Claude Shannon (1948): maximum entropy in frequentist language
- Edwin Thompson Jaynes: Bayesian re-interpretation
- Harold Jeffreys (1939): standard work on (Bayesian) probability
- Richard T. Cox (1946): derivation of probability laws
- Computers: flowering of Bayesian methods
Overview

1. Origins of probability
2. Frequentist methods and statistics
3. Principles of Bayesian probability theory
4. Monte Carlo computational methods
5. Applications
   - Classification
   - Regression analysis
6. Conclusions and references
Probability = frequency

Straightforward:
- Number of 1s in 60 dice throws \( \approx 10 \): \( p = 1/6 \)
- Probability of plasma disruption \( p \approx N_{\text{disr}} / N_{\text{tot}} \)

Less straightforward:
- Probability of fusion electricity by 2050?
- Probability of mass of Saturn 90 \( m_A \leq m_S < 100 m_A \)?
Populations vs. sample

- **PARAMETER**
  - Population mean ($\mu$)
  - Population standard deviation ($\sigma$)

- **STATISTIC**
  - Sample mean ($\bar{x}$)
  - Sample standard deviation ($s$)

We want to know about these
We have these to work with

(Random) Selection

Population

Sample
E.g. weight \( w \) of Belgian men: unknown but \textit{fixed} for every individual

- Average weight \( \mu_w \) in population?

- \textit{Random variable} \( W \)

- Sample: \( W_1, W_2, \ldots, W_n \)

- Average weight: \textit{statistic} (estimator) \( \overline{W} \)

- Central limit theorem:

\[
W \sim p(W|\mu_w, \sigma_w) \Rightarrow \overline{W} \sim \mathcal{N}(\mu_w, \sigma_w / \sqrt{n})
\]
Maximum likelihood (ML) principle:

\[
\hat{\mu}_w = \arg \max_{\mu_w \in \mathbb{R}^+} p(W_1, \ldots, W_n | \mu_w, \sigma_w)
\]

\[
\approx \arg \max_{\mu_w \in \mathbb{R}^+} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_w}} \exp \left[ -\frac{(W_i - \mu_w)^2}{2\sigma_w^2} \right]
\]

\[
= \arg \max_{\mu_w \in \mathbb{R}^+} \frac{1}{\sqrt{2\pi\sigma_w}} \exp \left[ -\sum_{i=1}^{n} \frac{(W_i - \mu_w)^2}{2\sigma_w^2} \right]
\]

ML estimator (known \(\sigma_w\)):

\[
\hat{\mu}_w = \bar{W} = \frac{1}{n} \sum_{i=1}^{n} W_i
\]
Frequentist hypothesis testing

- Weight of Dutch men compared to Belgian men (populations)
- Observed sample averages $\overline{W}_{NL}, \overline{W}_{BE}$
- **Null hypothesis** $H_0$: $\mu_{w,NL} = \mu_{w,BE}$
- Test statistic:
  $$\frac{\overline{W}_{NL} - \overline{W}_{BE}}{\sigma_{W_{CZ}-W_{BE}}} \sim \mathcal{N}(0, 1) \quad \text{(under } H_0)$$

![Normal distribution with z = 1.96 and p = 0.05]
Overview

1. Origins of probability
2. Frequentist methods and statistics
3. Principles of Bayesian probability theory
4. Monte Carlo computational methods
5. Applications
   - Classification
   - Regression analysis
6. Conclusions and references
Statistical (aleatoric) vs. epistemic uncertainty?

Every piece of information has uncertainty

Uncertainty = lack of information

Observation may reduce uncertainty

Probability (distribution) quantifies uncertainty
Example: physical sciences

- Measurement of physical quantity

- Origin of stochasticity:
  - Apparatus
  - Microscopic fluctuations

- Systematic uncertainty is assigned a probability distribution

- E.g. coin tossing, voltage measurement, probability of hypothesis vs. another, ...

- Bayesian: no random variables
What is probability?

- **Objective Bayesian view**
- **Probability** = real number $\in [0, 1]
- Always conditioned on known information
- **Notation:**
  \[ p(A|B) \] or \[ p(A|I) \]
- **Extension of logic:** measure of degree to which $B$ implies $A$
- **Degree of plausibility**, but subject to consistency
- **Same information** $\Rightarrow$ **same probabilities**
- **Probability distribution:** outcome $\rightarrow$ probability
Joint, marginal and conditional distributions

$p(x, y), p(x), p(y), p(x|y), p(y|x)$
Example: normal distribution

- Normal/Gaussian _probability density function_ (PDF):

\[
p(x|\mu, \sigma, I) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]
\]

- Probability \( x_1 \leq x < x_1 + dx \)
- Inverse problem: \( \mu, \sigma \) given \( x \)?
Bayes’ theorem

\[ p(\theta|x, I) = \frac{p(x|\theta, I)p(\theta|I)}{p(x|I)} \]

- **Likelihood**: misfit between model and data
- **Prior** distribution: ‘expert’ or diffuse knowledge
- **Evidence**:
  \[ p(x|I) = \int p(x, \theta|I) \, d\theta = \int p(x|\theta, I)p(\theta|I) \, d\theta \]
- **Posterior** distribution
Bayes’ theorem

\[ p(\theta|x, I) = \frac{p(x|\theta, I)p(\theta|I)}{p(x|I)} \]

- **Likelihood**: misfit between model and data
- **Prior** distribution: ‘expert’ or diffuse knowledge
- **Evidence**:
  \[ p(x|I) = \int p(x, \theta|I) \, d\theta = \int p(x|\theta, I)p(\theta|I) \, d\theta \]
- **Posterior** distribution
Bayes’ theorem

\[ p(\theta|x, I) = \frac{p(x|\theta, I)p(\theta|I)}{p(x|I)} \]

- **Likelihood**: misfit between model and data
- **Prior** distribution: ‘expert’ or diffuse knowledge
- **Evidence**:
  \[ p(x|I) = \int p(x, \theta|I) \, d\theta = \int p(x|\theta, I)p(\theta|I) \, d\theta \]
- **Posterior** distribution
Bayes’ theorem

\[ p(\theta|x, I) = \frac{p(x|\theta, I)p(\theta|I)}{p(x|I)} \]

- **Likelihood**: misfit between model and data
- **Prior** distribution: ‘expert’ or diffuse knowledge
- **Evidence**:
  \[ p(x|I) = \int p(x, \theta|I) \, d\theta = \int p(x|\theta, I)p(\theta|I) \, d\theta \]
- **Posterior** distribution

\( x = \text{data vector} \)
\( \theta = \text{vector of model parameters} \)
\( I = \text{implicit knowledge} \)
Practical considerations

- Updating state of knowledge = learning
- ‘Uninformative’ prior (distribution)
- Assigning uninformative priors:
  - Transformation invariance (Jaynes ’68): e.g. uniform for location variable
  - Principle of indifference
  - Testable information: maximum entropy
  - …
Mean of a normal distribution: uniform prior

- \( n \) measurements \( x_i \rightarrow x \)
- Independent and identically distributed \( x_i \):

\[
p(x|\mu, \sigma, I) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp \left[ -\frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

- Bayes’ rule:

\[
p(\mu, \sigma|x, I) \propto p(x|\mu, \sigma, I)p(\mu, \sigma|I)
\]

- Suppose \( \sigma \equiv \sigma_e \rightarrow \) delta function
- Assume \( \mu \in [\mu_{\text{min}}, \mu_{\text{max}}] \rightarrow \) uniform prior:

\[
p(\mu|I) = \begin{cases} 
1, & \text{if } \mu \in [\mu_{\text{min}}, \mu_{\text{max}}] \\
\mu_{\text{max}} - \mu_{\text{min}}, & \text{otherwise}
\end{cases}
\]

- Let \( \mu_{\text{min}} \rightarrow -\infty, \mu_{\text{max}} \rightarrow +\infty \Rightarrow \text{improper prior} \)
- Ensure proper posterior
Posterior for $\mu$

- Posterior:
  \[
p(\mu|x, I) \propto \exp \left[ -\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma_e^2} \right]
  \]

- Define
  \[
  \bar{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i, \quad (\Delta x)^2 \equiv \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
  \]

- Adding and subtracting $2n\bar{x}^2$ (‘completing the square’),
  \[
p(\mu|x, I) \propto \exp \left\{ -\frac{1}{2\sigma_e^2/n} \left[ (\mu - \bar{x})^2 + (\Delta x)^2 \right] \right\}
  \]

- Retaining dependence on $\mu$,
  \[
p(\mu|x, I) \propto \exp \left[ -\frac{(\mu - \bar{x})^2}{2\sigma_e^2/n} \right]
  \]

- $\mu \sim N(\bar{x}, \sigma_e^2/n)$
Mean of a normal distribution: normal prior

- Normal prior: \( \mu \sim \mathcal{N}(\mu_0, \tau^2) \)
- Posterior:

\[
p(\mu|x,I) \propto \exp \left[ -\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2_e} \right] \times \exp \left[ -\frac{(\mu - \mu_0)^2}{2\tau^2} \right]
\]

- Expanding and completing the square,

\[
\mu \sim \mathcal{N}(\mu_n, \sigma_n^2),
\]

where

\[
\mu_n \equiv \sigma_n^2 \left( \frac{n}{\sigma^2_e} \bar{x} + \frac{1}{\tau^2} \mu_0 \right) \quad \text{and} \quad \mu_n \equiv \sigma_n^2 \left( \frac{n}{\sigma^2_e} + \frac{1}{\tau^2} \right)^{-1}
\]

- \( \mu_n \) is weighted average of \( \mu_0 \) and \( \bar{x} \)
Intrinsic part of Bayesian theory, enabling coherent learning

High-quality data may be scarce

Regularization of ill-posed problems (e.g. tomography)

Prior can retain influence: e.g. regression with errors in all variables
Repeated measurements $\rightarrow$ information on $\sigma$

Scale variable $\sigma \rightarrow$ **Jeffreys' scale prior**:

$$p(\sigma | I) \propto \frac{1}{\sigma}, \quad \sigma \in ]0, +\infty[$$

Posterior:

$$p(\mu, \sigma | x, I) \propto \frac{1}{\sigma^n} \exp \left[ - \frac{(\mu - \bar{x})^2 + (\Delta x)^2}{2\sigma^2 / n} \right] \times \frac{1}{\sigma}$$
Marginalization = integrating out a (nuisance) parameter:

\[ p(\mu | x, I) = \int_0^{+\infty} p(\mu, \sigma | x, I) \, d\sigma \]

\[ \propto \int_0^{+\infty} \frac{1}{2} \left[ \frac{(\mu - \bar{x})^2 + (\Delta x)^2}{2/n} \right]^{-n/2} s^{n-1} e^{-s} \, ds \]

\[ = \frac{1}{2} \Gamma \left( \frac{n}{2} \right) \left[ \frac{(\mu - \bar{x})^2 + (\Delta x)^2}{2/n} \right]^{-n/2} \]

where

\[ s \equiv \frac{(\mu - \bar{x})^2 + (\Delta x)^2}{2\sigma^2/n} \]
Marginal posterior for $\mu$ (2)

- After normalization:

$$p(\mu|x, I) = \frac{\Gamma \left( \frac{n}{2} \right)}{\sqrt{\pi (\Delta x)^2 \Gamma \left( \frac{n-1}{2} \right)}} \left[ 1 + \frac{(\mu - \bar{x})^2}{(\Delta x)^2} \right]^{\frac{-n}{2}}$$

- Changing variables,

$$t \equiv \frac{(\mu - \bar{x})^2}{\sqrt{(\Delta x)^2 / (n-1)}}$$

$$p(t|x, I) dt \equiv p(\mu|x, I) d\mu,$$

$$p(t|x, I) = \frac{\Gamma \left( \frac{n}{2} \right)}{\sqrt{(n-1)\pi \Gamma \left( \frac{n-1}{2} \right)}} \left[ 1 + \frac{t^2}{n-1} \right]^{\frac{-n}{2}}$$

- Student’s $t$-distribution with parameter $\nu = n - 1$

- If $n \gg 1$,

$$p(\mu|x, I) \rightarrow \frac{1}{\sqrt{2\pi(\Delta x)^2 / n}} \exp \left[ -\frac{(\mu - \bar{x})^2}{2(\Delta x)^2 / n} \right]$$
Marginal posterior for $\sigma$

- Marginalization of $\mu$:

$$p(\sigma|x,I) = \int_{-\infty}^{+\infty} p(\mu, \sigma|x,I) \, d\mu$$

$$\propto \frac{1}{\sigma^n} \exp \left[-\frac{(\Delta x)^2}{2\sigma^2/n}\right]$$

- Setting $X \equiv n(\Delta x)^2/\sigma^2$,

$$p(X|x,I) = \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} X^{\frac{k}{2}-1} e^{-\frac{X}{2}}, \quad k \equiv n - 1$$

- $\chi^2$ distribution with parameter $k$
The Laplace approximation (1)

- Laplace (saddle point) approximation of distributions around the mode (= maximum)
- E.g. marginal for $\mu$:

$$p(\mu|x, I) = \frac{\Gamma \left( \frac{n}{2} \right)}{\sqrt{\pi(\Delta x)^2 \Gamma \left( \frac{n-1}{2} \right)}} \left[ 1 + \frac{(\mu - \bar{x})^2}{(\Delta x)^2} \right]^{-\frac{n}{2}}$$

- Taylor expansion around mode:

$$\ln[p(\mu|x, I)] \approx \ln[p(\bar{x}|x, I)] + \frac{1}{2} \frac{d^2(\ln p)}{d\mu^2} \bigg|_{\mu=\bar{x}} (\mu - \bar{x})^2$$

$$= \ln \left[ \Gamma \left( \frac{n}{2} \right) \right] - \ln \left[ \Gamma \left( \frac{n-1}{2} \right) \right] - \frac{1}{2} \ln \left[ \pi(\Delta x)^2 \right] - \frac{n}{2(\Delta x)^2} (\mu - \bar{x})^2$$
The Laplace approximation (2)

On the original scale:

\[ p(\mu|x, I) \approx \frac{\Gamma \left( \frac{n}{2} \right)}{\Gamma \left( \frac{n-1}{2} \right)} \frac{1}{\sqrt{\pi (\Delta x)^2}} \exp \left[ - \frac{(\mu - \bar{x})^2}{2(\Delta x)^2 / n} \right] \]

Standard deviation \( \sigma_L \) \( \rightarrow \) curvature of \( \ln p \):

\[ \sigma_L = \left. \left[ - \frac{d^2 (\ln p)}{d\mu^2} \right] \right|_{\mu = \bar{x}} \right]^{-1/2} \]
Laplace approximation: example 1
Laplace approximation: example 2
Multivariate Laplace approximation

- For $\theta = [\theta_1, \ldots, \theta_p]^t$,

  $$p(\theta|\theta_0, I) \propto \exp \left[ \frac{1}{2} (\theta - \theta_0)^t \left[ \nabla \nabla (\ln p) \right]_{\theta=\theta_0} (\theta - \theta_0) \right]$$

- $\nabla \nabla (\ln p)$: Hessian matrix, where

  $$\Sigma_L = - \left\{ \left[ \nabla \nabla (\ln p) \right]_{\theta=\theta_0} \right\}^{-1}$$
Model comparison (hypothesis testing)

- Let \( \{H_i\} \) be complete set of hypotheses
- Data \( D \) to support or reject hypotheses
- Bayes’ rule:
  \[
p(H_i|D,I) = \frac{p(D|H_i,I)p(H_i|I)}{p(D|I)}, \quad p(D|I) = \sum_i p(D|H_i,I)p(H_i|I)
\]
- Assume single hypothesis \( H \) and complement \( \overline{H} \)
- **Odds ratio** \( o \):
  \[
o \equiv \frac{p(H|D,I)}{p(\overline{H}|D,I)} = \frac{p(D|H,I)}{p(D|\overline{H},I)} \frac{p(H|I)}{p(\overline{H}|I)}
\]
  Bayes factor \quad Prior odds
E.g. $n$ measurements $x_i$ of a quantity $x$

Assume normal distribution with known variance $\sigma^2$

Question: are the data compatible with mean $\mu = \mu_0$?

- Yes: $H$
- No: $\overline{H}$

Under $H$:

$$p(\bar{x}|H, I) = C \exp \left[ -\frac{1}{2\sigma^2/n} (\bar{x} - \mu_0)^2 \right]$$

Under $\overline{H}$:

$$p(\bar{x}|\overline{H}, I) = \int p(\bar{x} | \mu, \sigma, I) p(\mu | \overline{H}, I) \, d\mu$$

(1)
Assume bounds $\mu_{\text{min}}$ and $\mu_{\text{max}}$:

$$p(\mu|H,I) = \frac{1}{|\mu_{\text{max}} - \mu_{\text{min}}|} 1(\mu_{\text{min}} \leq \mu \leq \mu_{\text{max}})$$

Then (1) becomes

$$p(\bar{x}|H,I) = \frac{C}{|\mu_{\text{max}} - \mu_{\text{min}}|} \int_{\mu_{\text{min}}}^{\mu_{\text{max}}} \exp \left[ -\frac{1}{2\sigma^2/n} (\bar{x} - \mu)^2 \right] \, d\mu$$

With $\text{SE} \equiv \sigma / \sqrt{n} =$ standard error, assume

$$|\bar{x} - \mu_{\text{min}}|, |\bar{x} - \mu_{\text{max}}| \gg \text{SE}$$
Testing a Gaussian mean (3)

- Then

\[ p(\bar{x}|H, I) \approx \frac{C}{|\mu_{\text{max}} - \mu_{\text{min}}|} \int_{-\infty}^{+\infty} \exp \left[- \frac{1}{2\sigma^2/n} (\bar{x} - \mu)^2 \right] d\mu \]

\[ = \frac{C \ SE \ \sqrt{2\pi}}{|\mu_{\text{max}} - \mu_{\text{min}}|} \]

- Bayes factor BF:

\[ BF = \frac{p(\bar{x}|H, I)}{p(\bar{x}|H, I)} = \frac{|\mu_{\text{max}} - \mu_{\text{min}}|}{SE} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \]

\[ z \equiv \frac{\bar{x} - \mu_0}{SE} \]

- Cf. frequentist hypothesis test
Overview

1. Origins of probability
2. Frequentist methods and statistics
3. Principles of Bayesian probability theory
4. Monte Carlo computational methods
5. Applications
   - Classification
   - Regression analysis
6. Conclusions and references
Monte Carlo sampling

- (Strong) nonlinearities through prior or likelihood
- Skewed, heavy-tailed, multimodal
- Sampling (simulating) from distributions $p$
- Estimate properties of $p$
- Marginalization: sampling from joint distribution → sampling from marginals
Markov chain Monte Carlo (MCMC) (1)

- **Metropolis-Hastings** sampling, **Gibbs** sampling

- Simple algorithms, but implementation issues might arise

- **Target distribution** $\pi(\theta|I)$

- **Markov chain** $\theta^{(t)}$: conditional on last state

- Sample from **proposal distribution**
Subchains sample from corresponding marginals

Monte Carlo averages, e.g.

\[ \bar{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \theta_j^{(t_c+i)}, \quad (\Delta \theta_j)^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \theta_j^{(t_c+i)} - \bar{\theta}_j \right)^2 \]
1. At $t = 0$, start from $\theta^{(t)} = \begin{bmatrix} \theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_p^{(t)} \end{bmatrix}^t$

2. repeat

3. for $j = 1$ to $p$

   
   \[ y \sim q_j \left( \eta \bigg| \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, \theta_j^{(t)}, \theta_{j+1}^{(t)}, \ldots, \theta_p^{(t)}, \mathbf{I} \right) \]

   \[ \theta_j^{(t+1)} = \begin{cases} 
   y, & \text{with probability } \rho \\
   \theta_j^{(t)}, & \text{with probability } 1 - \rho 
   \end{cases} \]

4. end
Metropolis-Hastings algorithm (2)

- **Acceptance probability:**

\[
\rho \left( y, \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, \theta_j^{(t)}, \theta_{j+1}^{(t)}, \ldots, \theta_p^{(t)} \right) \\
\equiv \min \left\{ 1, \frac{\pi \left( \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, y, \theta_{j+1}^{(t)}, \ldots, \theta_p^{(t)}, I \right)}{\pi \left( \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, \theta_j^{(t+1)}, \theta_{j+1}^{(t)}, \ldots, \theta_p^{(t)}, I \right)} \times \frac{q_j \left( \theta_j^{(t)} \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, y, \theta_{j+1}^{(t)}, \ldots, \theta_p^{(t)}, I \right)}{q_j \left( \theta_j^{(t)} \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, \theta_j^{(t)}, \theta_{j+1}^{(t)}, \ldots, \theta_p^{(t)}, I \right)} \right\}
\]

- **Example:**

\[
q_j \left( \eta \mid \theta_j^{(t+1)}, I \right) \propto \exp \left[ -\frac{\left( \eta - \theta_j^{(t)} \right)^2}{2\sigma_j^2} \right]
\]
MCMC example
Overview

1. Origins of probability
2. Frequentist methods and statistics
3. Principles of Bayesian probability theory
4. Monte Carlo computational methods
5. Applications
   - Classification
   - Regression analysis
6. Conclusions and references
Overview

1. Origins of probability
2. Frequentist methods and statistics
3. Principles of Bayesian probability theory
4. Monte Carlo computational methods
5. Applications
   - Classification
   - Regression analysis
6. Conclusions and references
Classification or clustering

Binary classification:

Multi-class classification:
Simple Bayesian classification

- $M$ clusters of in total $n$ data points $x_i$ in $P$-dimensional space
- Known class labels $\omega_j$ ($j = 1, \ldots, M$) of $x_i$
- Bayes’ rule for new point $x$:

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

- Maximum a posteriori (MAP) classification rule for $x$:

Assign $x$ to $\omega_i = \arg \max_{\omega_j} p(\omega_j|x, I) = \arg \max_{\omega_j} p(x|\omega_j, I)p(\omega_j|I)$
Examples of priors and likelihoods

- **Examples of prior probabilities (indifference):**
  - \( p(\omega_i|I) = p(\omega_j|I), \forall i, j \)
  - Count class membership:
    \[
    p(\omega_i|I) \equiv \frac{n_i}{n}, \quad i = 1, \ldots, M
    \]

- **Examples of likelihoods:**
  - **Naive Bayesian classifier:**
    \[
    p(x|\omega_i) = \prod_{k=1}^{p} p(x_k|\omega_i), \quad i = 1, \ldots, M
    \]
  - Multivariate Gaussian:
    \[
    p(x|\omega_j, I) = \frac{1}{(2\pi)^{p/2}|\Sigma_j|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu_j)^t \Sigma_j^{-1} (x - \mu_j) \right]
    \]
Optimality of Bayesian classifier

Bayesian classifier minimizes probability of misclassification
Bayesian classifier minimizes probability of misclassification.
MAP classification: decision surfaces

- **MAP**: maximize w.r.t. $\omega_j$:
  \[
  \ln p(\omega_j | x, I) = \ln p(x | \omega_j, I) + \ln p(\omega_j | I)
  \]

- Define ($M = 2$):
  \[
  g(x) \equiv \ln p(\omega_1 | x, I) - \ln p(\omega_2 | x, I)
  \]

- For normal likelihood:
  \[
  g(x) = \frac{1}{2} \left( x^t \Sigma_2^{-1} x - x^t \Sigma_1^{-1} x \right) + \mu_1^t \Sigma_1^{-1} x - \mu_2^t \Sigma_2^{-1} x
  \]

- Constant:
  \[
  -\frac{1}{2} \mu_1^t \Sigma_1^{-1} \mu_1 + \frac{1}{2} \mu_2^t \Sigma_2^{-1} \mu_2 + \frac{1}{2} \ln \frac{\Sigma_2}{\Sigma_1} + \ln \frac{p(\omega_1 | I)}{p(\omega_2 | I)}
  \]

- $g(x)$ separates classes: **decision hypersurface**
Discriminant analysis
Quadratic discriminant analysis (QDA)

Linear discriminant analysis (LDA): $\Sigma_1 = \Sigma_2$
ELM type classification

\[ P_{\text{input}} - 1.41 \Gamma_{D_2} = 7.47 \]

A. Shabbir et al., Fusion Eng. Des. 123, 717, 2017
Minimum distance classifiers (1)

- Linear classifier with normal likelihood ($\Sigma_1 = \Sigma_2 \equiv \Sigma$):

  $$g(x) = \theta^t(x - x_0) = 0,$$
  $$\theta \equiv \Sigma^{-1}(\mu_1 - \mu_2)$$
  $$x_0 \equiv \frac{1}{2}(\mu_1 + \mu_2) - \ln \frac{p(\omega_1|I)}{p(\omega_2|I)} \frac{\mu_1 - \mu_2}{\|\mu_1 - \mu_2\|_{\Sigma^{-1}}}$$
  $$\|\mu_1 - \mu_2\|_{\Sigma^{-1}} \equiv \sqrt{(\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)}$$

- Mahalanobis distance

- $p(\omega_1|I) = p(\omega_2|I) \Rightarrow$ Minimum Mahalanobis distance classifier (to cluster centroid)

- $\Sigma = \sigma^2 I \Rightarrow$ Minimum Euclidean distance classifier

- $k$-nearest neighbor classification ($k$NN)
Minimum distance classifiers (2)
Overview

1. Origins of probability
2. Frequentist methods and statistics
3. Principles of Bayesian probability theory
4. Monte Carlo computational methods
5. Applications
   - Classification
   - Regression analysis
6. Conclusions and references
Multilinear regression model

- Regression model:
  \[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon \]
  \[ \epsilon \sim \mathcal{N}(0, \sigma^2), \sigma \text{ known} \]

- Negligible uncertainty on \( x_j \)

- Likelihood:
  \[ p(y|x, \beta, \sigma, I) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left[ -\frac{1}{2\sigma^2} \left( y - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 \right], \]
  \[ \beta \equiv [\beta_0, \beta_p^t]^t, \quad \beta_p \equiv [\beta_1, \ldots, \beta_p]^t \]
Maximum likelihood solution

- Take \( n \) measurements:

\[
y = [y_1, \ldots, y_n]^t
\]

\[
X \equiv \begin{bmatrix}
1 & x_{11} & \cdots & x_{1p} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & \cdots & x_{np}
\end{bmatrix}
\]

- Conditional independence:

\[
p(y|X, \beta, \sigma, I) = (2\pi)^{-n/2}\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^t(y - X\beta)\right]
\]

- ML: maximize w.r.t. \( \beta \):

\[
0 = \nabla_\beta (y - X\beta)^t(y - X\beta) = -2X^ty + 2X^tX\beta
\]

\[
\Rightarrow \quad \beta_{ML} = (X^tX)^{-1}X^ty = \beta_{LS}
\]

Moore-Penrose pseudoinverse
MAP solution and posterior

- Uniform priors on $\beta_j$ (not the most uninformative!):
  
  $$p(\beta|y, X, \sigma, I) \propto \exp \left[ -\frac{1}{2\sigma^2} (y - X\beta)^t(y - X\beta) \right]$$

- Due to linearity and Gaussianity: $\beta_{MAP} = \beta_{ML} = \beta_{LS}$

- Taylor expansion (exact!):
  
  $$\begin{align*}
  (y - X\beta)^t(y - X\beta) &= (y - X\beta_{MAP})^t(y - X\beta_{MAP}) \\
  &+ \frac{1}{2}(\beta - \beta_{MAP})^t 2X^tX(\beta - \beta_{MAP})
  \end{align*}$$

- Hence,
  
  $$p(\beta|y, X, \sigma, I) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \times \exp \left[ -\frac{1}{2}(\beta - \beta_{MAP})^t \Sigma^{-1}(\beta - \beta_{MAP}) \right] ,$$

  $$\Sigma \equiv \sigma^2(X^tX)^{-1}$$
New predictions by the model?

Posterior predictive distribution:

\[ p(y_{\text{new}}|x_{\text{new}}, y, X, \sigma, I) = \int_{\mathbb{R}^{p+1}} p(y_{\text{new}}, \beta|x_{\text{new}}, y, X, \sigma, I) \, d\beta = \int_{\mathbb{R}^{p+1}} p(y_{\text{new}}|x_{\text{new}}, \beta, I) \, p(\beta|y, X, \sigma, I) \, d\beta \]

But

\[ p(y_{\text{new}}|x_{\text{new}}, \beta, I) = \delta(y_{\text{new}} - \beta^t x_{\text{new}}) \]

Fix \( \beta_0 = y_{\text{new}} - \beta_1 x_{\text{new},1} - \cdots - \beta_p x_{\text{new},p} \)
Marginalize over $\beta_p$ with flat priors:

$$p(y_{new}|x_{new}, y, X, \sigma, I) \propto \sigma^{-n} \int_{\mathbb{R}^p} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[ y_i - y_{new} + \sum_{j=1}^{p} \beta_j (x_{new,j} - x_{ij}) \right]^2 \right\} d\beta_p$$

After (quite some) algebra, one finds simply

$$y_{new,MAP} = \sum_{j=1}^{p} x_{new,j} \beta_{MAP,j} + \beta_{MAP,0}$$

Simpler derivation based on properties of $\mathbb{E}$ and Var

General posterior more complicated!
Overview

1. Origins of probability
2. Frequentist methods and statistics
3. Principles of Bayesian probability theory
4. Monte Carlo computational methods
5. Applications
   - Classification
   - Regression analysis
6. Conclusions and references
Conclusions

- Frequentist vs. Bayesian methods: interpretation of probability

- Bayesian probability: extension of logic to situations with uncertainty

- Posterior probability of parameters or hypotheses

- Numerical approach in general

- Underlies or explains many machine learning techniques
References

- S.B. McGrayne, *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*, Yale University Press, 2011