# ST451 - Lent term
# Bayesian Machine Learning

Kostas Kalogeropoulos

Bayesian Inference Concepts

# Outline

# Outline

1. Practical Information - Course content
2. Machine Learning and Bayesian Inference
3. Bayes Estimators, Credible Intervals and Forecasting
4. Bayesian Inference via Monte Carlo methods

# Teaching

My name is Kostas Kalogeropoulos and will be doing the lectures and one computer class group.

Phil Chan and Gianluca Giuduce will teach the other two computer classes groups.

Lectures: 2 hours every week on Monday 13:00–15:00 in room NAB.2.04 (except week 6 which is NAB.LG.01).

Computer Classes: 3 groups Mon 15:00-16:30, Tue 16:00-17:30 and Thu 15:00-16:30.

# In Class

During lectures we will cover the theory and go through several examples.

Recordings will be available on Moodle but try to attend lectures anyway.

During computer classes we will go through the computer part of the course.

Class attendance is compulsory and will be recorded on LSE for You.

# Moodle and Course Webpage

All relevant material will be posted on either Moodle or the course webpage. The Moodle Enrolment key: Laplace

- Slides for each lecture.
- Recording of lectures.
- Code for each computer class
- Problem sets each Monday.
- Solution of problem sets with code once you have handed them back.

# Computing

- Python will be used throughout the course.
- You can either bring your laptop to the computer classes or use the room's PC.
- Install Anaconda from the link below (Python 3.7 version)

  `https://www.anaconda.com/download/`

# Weekly formative assignments

Each week you will be assigned a problem set containing both theoretical and computer exercises.

It will be due next week. Submit in Columbia House Box 34. Write your class group number in the first page

Hand in everything you are able to solve even if it is not complete. Marks don't count in the final grade but are recorded on LSE for You.

Problem sets will be returned marked with feedback during the class of the week after you handed them in.

# Assessment

An individual project will be assigned on week 7 and will be due Tuesday, May 12th noon. You will be required to analyse data of your choice using the taught Bayesian Machine Learning techniques and present your findings through a paper-like report.

During summer term the course is assessed by a 2 hour written exam.

The final grade will be determined by the above with equal weights (50-50%).

# Questions and Feedback

1. Ask questions in class.

2. Office and Feedback hours: COL.6.10 on Monday 10:30–12:00.

3. Use the forum on Moodle.

4. Feel free to send emails but please try to avoid questions that might be of interest for all the class; use Moodle forum instead.

# Syllabus

- Weeks 1-2: Bayesian Inference Concepts
- Weeks 2-3: Linear Regression
- Week 4: Classification
- Week 5: Graphical Models
- Week 6: Mixture Models and Clustering
- Week 7: Approximate Inference
- Week 8: Sampling methods
- Week 9: Sequential Data
- Week 10: Gaussian Processes

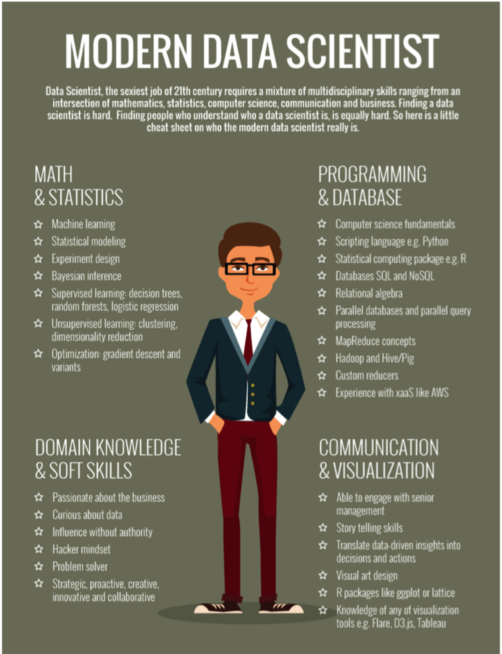
# Reading

Lecture slides will be sufficient for exam purposes but for further reading you can check the books below:

- C. M. Bishop, Pattern Recognition and Machine Learning, Springer 2006
- K. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012
- D. Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press 2012
- S. Rogers and M. Girolami, A First Course in Machine Learning, Second Edition, Chapman and Hall/CRC, 2016

# Outline

# Skills of a Data Scientist



# Machine Learning and related fields

**Machine Learning:** A set of methods that can automatically detect patterns in data, and then use them to predict future data, or take decisions under uncertainty.

**Related fields:**

- *Data Mining:* Data Mining focuses on discovering unknown patterns in the data. Machine Learning wants to use these patterns to complete some tasks.
- *Optimisation:* Optimisation aims in minimising some kind of loss in the data. Machine Learning focuses on unseen future data.
- *Statistics:* Closely linked. In statistics we typically have probability models. In machine we may also have algorithmic models.

# Why Bayesian vs Standard Machine Learning?

- **Being Bayesian without realising it:** Penalised methods, such as Lasso/Ridge regression, make more sense as Bayesian methods.
- **Know when you don't know:** Natural framework for handling uncertainty in predictions.
- **Principled probabilistic framework:** Leads to techniques such as graphical models, sequential methods, Gaussian processes.
- **Philosophical background:** Natural framework for learning.

# Defining probability

**Frequentist definition:** If the experiment was repeated many times, probability of $A$ is the frequency $f_n$ in which a given event $A$ is realised.

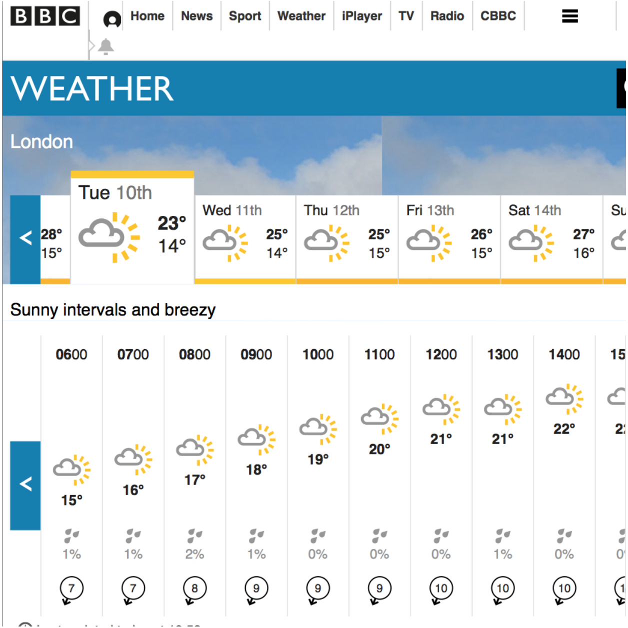
$$P(A) = \lim_{n\to\infty} f_n.$$

**Subjective definition:** $P(A)$: a number in $[0, 1]$ reflecting our beliefs on how likely $A$ is (0:impossible, 1:certain).

**Example:** Probability of heads when tossing a fair coin?

- If we toss a fair coin $n$ times, where $n$ is large, count the number of times we get head say $k_n$ over $n$. If the coin is fair, we should get 0.5.
- Alternatively, we can argue that it makes sense for this probability to be 0.5 as we believe that heads and tails are equally likely.

# Probability of rain in a particular day?



# Frequentist Probability and Time Travel?

# Frequentist Probability and Multiverse?

# Statistical Analysis or Machine learning setting

- Consider data $y = (y_1, y_2 \ldots, y_n)$ from a real world application.
- Assign a suitable probability model aka likelihood for data $y$ and parameter(s) $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$.

$$f(y_1, \ldots, y_n | \theta_1, \ldots, \theta_p) = f(y|\theta).$$

- Use $y$ to learn about $\theta$ and answer the relevant questions or predict future $y$.

# Prior Information

Consider the following 3 experiments where the probability of a correct answer $\theta$ is of interest.

1. A lady claims that by tasting a cup of tea with milk she can tell whether the milk was poured into the cup before the tea. In 9 out of 10 trials she gets it right.
2. A musical expert claims that he can distinguish by a small music part whether it is Mozart or Haydn. He gets it right in 9 out of 10 times.
3. A drunk man claims he can predict the outcome of a fair coin flip. In 9 flips out of 10 he is correct.

**Frequentist approach:** Test $H_0 : \theta = 0.5$ vs $H_1 : \theta > 0.5$.
This gives p-value $\approx$ 0.01 concluding genuine skill in all cases.

# Bayes theorem for events

Clearly there exists some relevant information prior to the experiments. Bayesian inference formally seeks to utilise prior information on $\theta$ by the so-called prior distribution $\pi(\theta)$

**Bayes Theorem for Events:** In terms of events and their probabilities, let $A$ and $B$ be two events such that $P(A) > 0$. then $P(B|A)$ and $P(A|B)$ are related by

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|B^c)P(B^c)}$$

More generally if $B_1$, $B_2$, ..., $B_k$ form a partition ($k$ can be $\infty$), we can write for all $j = 1, \ldots, k$

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)}, \text{ where } P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

# Bayesian Statistical model

We treat $\theta$ as a random variable and assign the prior pdf $\pi(\theta)$. The prior reflects our beliefs on $\theta$ before seeing the data.

The posterior distribution reflects our beliefs on $\theta$ after seeing the data and is the main object for inference. It is given by

$$\pi(\theta|y) \quad = \tfrac{f(y|\theta)\pi(\theta)}{f(y)}, \text{ where } f(y) = \int f(y|\theta)\pi(\theta)d\theta$$

The term $f(y)$ is known as marginal likelihood or evidence and reflects probability of the data under the adopted probability model. It may also be viewed as a normalising constant.

# Features of Bayesian Inference

- **Prior information:** Every problem is different and has its own context which can (in theory) be reflected via the prior distribution.

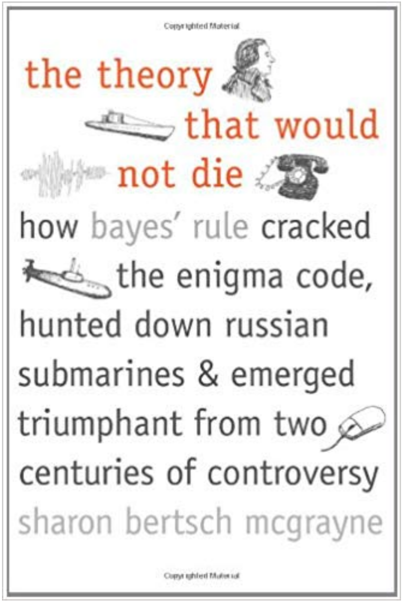
  Criticism: It is not always clear how to define the prior distribution. Results depend on the choice of it.

- **Subjective probability:** Accepting the subjective basis of knowledge. Also always defined contrary to the frequentist definition that applies to inherently repeatable events.

  Criticism: No guarantee that your quantification of uncertainty will be seen by others as ‘good’.

- **Common misconception:** Bayesians believe that there is a single unknown value for $\theta$. A distribution is assigned only to express subjective uncertainty not because the truth is random.

# History of Frequentist vs Bayesian Inference



# Outline

# Bayes Estimators

Point estimators: functions of $y$ (and other known things but not $\theta$), that provide an educated guess for $\theta$. e.g. MLEs, Method of Moments, least square estimators etc.

Bayes estimators provide another alternative. They minimise the posterior and Bayes risk (beyond the scope of the course).

We will focus on the following Bayes estimators:

- The posterior mean $\hat{\theta} = E(\theta|y) = \int_\theta \theta\pi(\theta|y)d\theta$.
- The posterior median $\hat{\theta} = q$ such that $\pi(\theta \leq q|y) = 0.5$.
- The posterior mode, aka MAP, $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \pi(\theta|y)$.

# Bayesian (Credible) Intervals

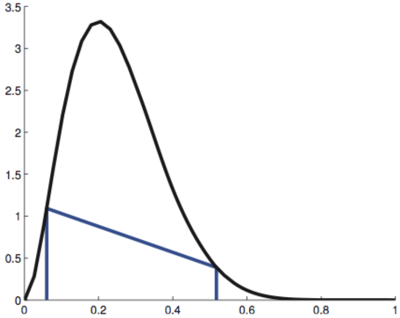
**Frequentist 95% Confidence Interval for $\theta$:** If the experiment was repeated many times and we constructed a confidence interval each time with the same procedure, 95% of these intervals would contain the true $\theta$.

**(Bayesian) 95% Credible Interval for $\theta$:.** $\theta$ is in a 95% credible interval with probability 95%.

Note that there exist many 95% credible intervals as well as many 95% confidence intervals.

**Construction:** A 95% credible interval, is usually defined by the 2.5% to the 97.5% points of $\pi(\theta|y)$. Another option is the Highest Posterior Density intervals.

# Illustration of Credible Intervals




(a) (b)

# Beta-Binomial Example

Suppose that $y$ is Binomial$(n, \theta)$. The likelihood is given by

$$f(y|\theta) = \left( \begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y} \propto \theta^y (1-\theta)^{n-y}$$

As $0 < \theta < 1$ a corresponding distribution must be chosen as the prior, e.g. the Beta$(\alpha, \beta)$, for some known positive $\alpha$, $\beta$.

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The posterior distribution can then be obtained as

$$\begin{aligned} \pi(\theta|y) \propto f(y|\theta)\pi(\theta) \propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ = \theta^{\alpha+y-1}(1-\theta)^{n-y+\beta-1} \\ \stackrel{\mathcal{D}}{=} \text{Beta}(\alpha+y, n-y+\beta) \end{aligned}$$

# Beta-Binomial Example (cont'd)

**Bayes Estimators:**

- Posterior mean, equal to $\frac{\alpha+y}{\alpha+\beta+n}$.
  Different than the MLE which is $y/n$
- Posterior mode, equal to $\frac{\alpha+y-1}{\alpha+\beta+n-2}$.
  Coincides with the MLE for $\alpha = \beta = 1$, i.e. Uniform$(0, 1)$ prior.

**Credible Intervals:** Use the 2.5-th and 97.5-th percentiles - todays computer class.

# Prior Specification

**Prior Elicitation:** Use existing information about $\theta$,
e.g. if for $\theta > 0$ it is known that $E[\theta] = 5$ and $\mathrm{Var}[\theta] = 4$, assign the Gamma$(6.25, 1.25)$.

What if no information is available?

**Transformation Invariance principle:** If no information is available for $\theta$ then no information should be available for any deterministic function of $\theta$ either.

**Jeffreys prior:** $\pi(\theta) \propto |I(\theta)|^{1/2}$, $I(\theta)$ is Fisher's information.
If $\theta = g(\phi)$ the Jeffreys prior for $\phi$ is $\pi(\phi) \propto |I(\phi)|^{1/2}$.

Usually not a proper distribution but posterior can be proper.

# Transformation invariance of Fisher's information

**Fisher's information for $\theta$:** Given $f(y|\theta)$, it is defined as

$$I(\theta) = E_Y\left[\left(\frac{\partial \log f(y|\theta)}{\partial\theta}\right)^2\right] = -E_Y\left(\frac{\partial^2 \log f(y|\theta)}{\partial\theta^2}\right).$$

**Lemma:** Let $\theta = g(\phi)$. Then if $\theta \sim \pi_\theta(\theta)$ the pdf of $\phi$ is

$$\pi_\phi(\phi) = \pi_\theta(g(\phi))\left|\frac{\partial g(\phi)}{\partial\phi}\right| = \pi_\theta(\theta)\left|\frac{\partial\theta}{\partial\phi}\right|$$

**Transformation invariance proof:** Let $\pi_\theta(\theta) \propto I(\theta)^{1/2}$ and take

$$\pi_\phi(\phi) = \pi_\theta(\theta)\left|\frac{\partial\theta}{\partial\phi}\right| \propto E_Y\left[\left(\frac{\partial \log f(y|\theta)}{\partial\theta}\right)^2\right]^{1/2}\left(\left|\frac{\partial\theta}{\partial\phi}\right|^2\right)^{1/2}$$

$$= E_Y\left[\left(\frac{\partial \log f(y|\theta)}{\partial\theta}\frac{\partial\theta}{\partial\phi}\right)^2\right]^{1/2} = E_Y\left[\left(\frac{\partial \log f(y|\phi)}{\partial\phi}\right)^2\right]^{1/2} = I(\phi)^{1/2}$$

# 'Low' informative priors

In practice, when there is no prior information, a low informative distribution, e.g. with high variance is selected.

This is usually ok for point/interval estimation.

It can be dangerous for some Hypothesis testing cases though (to be discussed later).

# Bayesian Prediction/Forecasting

Let $y_n$ denote a future observation. Under the assumption that $y_n$ comes from the same probability model as $y$, we are interested predicting its value.

Under Bayesian Prediction/Forecasting this is done via the *(posterior-)predictive* distribution that combines the uncertainty of the unknown parameters $\theta$ as well as the uncertainty of the future observation:

$$f(y_n|y) = \int f(y_n|\theta)\pi(\theta|y)d\theta.$$

The predictive distribution can be used in different ways (e.g. point prediction, interval prediction, etc) depending on the forecasting task at hand.

# Poisson-Gamma Example

Let $y = (y_1, \ldots, y_n)$ with $y_i$'s being independent and Poisson($\lambda$). The likelihood is given by the joint density of the sample

$$f(y|\lambda) = \prod_{i=1}^{n} \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} \propto \exp(-n\lambda)\lambda^{\sum y_i}$$

As $\lambda > 0$ assign the Gamma($\alpha, \beta$) as the prior for $\lambda$

$$\pi(\lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$$

The posterior can then be obtained as

$$\begin{aligned}
\pi(\lambda|y) \propto f(y|\lambda)\pi(\lambda) \propto \exp(-n\lambda)\lambda^{\sum y_i}\lambda^{\alpha-1}\exp(-\beta\lambda) \\
= \lambda^{\alpha+\sum y_i - 1}\exp(-n\lambda - \beta\lambda) \\
= \lambda^{\alpha+\sum y_i - 1}\exp\left[-\lambda(n+\beta)\right] \\
\overset{\mathcal{D}}{=} \text{Gamma}(\alpha + \sum y_i, n + \beta)
\end{aligned}$$

# Poisson-Gamma Example (cont'd)

**Bayes Estimator:**
Posterior mean, equal to

$$\frac{\alpha + n\bar{y}}{n+\beta} = ... = \left(1 - \frac{n}{n+\beta}\right)\frac{\alpha}{\beta} + \frac{n}{n+\beta}\bar{y},$$

which is a weighted average between the prior mean and $\bar{y}$ (MLE). As $n \to \infty$ the posterior mean converges to $\bar{y}$.

**Credible Intervals:** Use the 2.5-th and 97.5-th percentiles - todays computer class.

# Poisson-Gamma Example - Prior

**Prior Elicitation:**
If prior knowledge for $\lambda$ exists, it can be used to specify $\alpha$ and $\beta$, e.g. if $E[\lambda] = 5$ and $\mathrm{Var}[\lambda] = 4$, set $\alpha = 6.25$, $\beta = 1.25$.

**Low Informative prior:**
Set $\alpha = \beta = 0.001$, then $\mathrm{Var}[\lambda] = 1000$ - quite 'flat'.

**Jeffreys prior:**
Fisher's information $I(\lambda) = n/\lambda$ (see this week's exercises), hence $\pi(\lambda) \propto \lambda^{-1/2}$. Corresponds the Gamma$(1/2 + \sum y_i, n)$ as posterior.

# A trick for calculating integrals

Let $\lambda$ be Gamma$(A, B)$ for some positive $A, B$. It's pdf can be written as

$$f(\lambda|A, B) = \frac{B^A}{\Gamma(A)} \lambda^{A-1} \exp(-B\lambda)$$

for all positive $A$, $B$ and $\lambda$.

Not that $f(\lambda|A, B)$ is a pdf so it integrates to 1. Hence we can write (for all positive $A$, $B$ and $\lambda$)

$$\int_0^\infty \lambda^{A-1} \exp(-B\lambda) d\lambda = \frac{\Gamma(A)}{B^A}$$

We can do this for all known pdfs.

# Poisson-Gamma (posterior-)predictive distribution

Let now $y_n$ denote a future observation from the same Poisson($\lambda$) model. The predictive distribution for $y_n$ is

$$
\begin{aligned}
f(y_n|y) &= \int_0^\infty f(y_n|\lambda)\pi(\lambda|y)d\lambda \\
&= \int_0^\infty \frac{\exp(-\lambda)\lambda_n^y}{y_n!}\frac{(n+\beta)^{\alpha+\sum y_i}}{\Gamma(\alpha+\sum y_i)}\lambda^{\alpha+\sum y_i-1}\exp(-(n+\beta)\lambda)d\lambda \\
&= \frac{(n+\beta)^{\alpha+\sum y_i}}{y_n!\Gamma(\alpha+\sum y_i)}\int_0^\infty \lambda^{\alpha+y_n+\sum y_i-1}\exp(-(n+\beta+1)\lambda)d\lambda \\
&= \frac{(n+\beta)^{\alpha+\sum y_i}}{y_n!\Gamma(\alpha+\sum y_i)}\frac{\Gamma(\alpha+y_n+\sum y_i)}{(n+\beta+1)^{\alpha+y_n+\sum y_i}}
\end{aligned}
$$

for $y_n = 0, 1, \ldots$.

# Outline

# Why arthe posterior or the predictive needed?

So far we saw methods to find the posterior or the predictive distributions. Why do we need them?

- To find their mean, median or mode of the posterior - Bayes Estimation.
- To find the 2.5-th and 97.5-th percentiles of the posterior - Credible Intervals.
- Same for the predictive to obtain a prediction or a prediction interval.
- May even want to calculate probabilities for certain values values of future observations.

Essentially we only need to calculate expectations wrt these distributions. No need to know them, it suffices to be able to simulate from them.

# Monte Carlo Calculation of Expectation

## Monte Carlo

Let $F(x)$ be a probability distribution and $h(x)$ be a function such that $E_X(h(X)) < \infty$. Also let $x = (x_1, \ldots, x_n)$ be a sample from $F$. Then law of large numbers implies that as $n \to \infty$

$$E_X(h(X)) \to \frac{1}{n} \sum_{i=1}^{n} h(x_i)$$

## Implementation

Draw $x_1, \ldots, x_n$ from $F$ and calculate the integral using the sample mean. The error becomes arbitrarily small as $n$ increases.

# Simulating from distributions

- For most known distributions simulation is straightforward using a computer package (and of course Python).

- If we can simulate from the likelihood and the posterior we can also simulate from the predictive distribution in the following two steps:
  1. Simulate samples $\theta^*$ from the posterior $\pi(\theta|y)$.
  2. Simulate samples of potential future data $y_n$ from the likelihood $f(y|\theta^*)$.

- It is actually possible to simulate from the posterior distribution without fully knowing it. We will come back to on Week 8.

# Today's lecture - Reading

Murphy: 2.1-2.7, 5.2.1, 5.2.2, 6.6.1 and 6.6.2