Bayesian Machine Learning - Lecture 1

Guido Sanguinetti

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
gsanguin@inf.ed.ac.uk

February 23, 2015

Welcome

- Broad introduction to statistical machine learning concepts within the Bayesian probabilistic framework
- Focus on using statistics as a modelling tool, and algorithms for efficient inference
- Key objective: theoretical and practical familiarity with some fundamental ML methods
- Structure: four two hours lectures and one two hours lab each week
- Assessment: coinciding with the labs. NB: I believe PhD students have already demonstrated their ability to pass exams.

Main refs

- D. Barber, Bayesian Reasoning and Machine Learning, CUP 2010
- Some slides also taken from the teaching material attached to the book (thanks David!)
- Other good books: C.M. Bishop, Pattern Recognition and Machine Learning (Springer 2006); K. Murphy, Machine Learning - a probabilistic perspective (MIT Press 2012). Rasmussen and Williams, Gaussian Processes for Machine Learning (MIT Press 2007) for Lecture 3
- Lecture 4 (Active Learning and Bayesian Optimisation): B. Settles, Active learning literature survey, sections 2 and 3 and Brochu et al, `http://arxiv.org/abs/1012.2599`
- Wikipedia also has good pages for most of the material
- **IMPORTANT FACT**: these slides are not a book!!!!

Today's lecture

A few things worth considering

- Mobile traffic in 2013 $= 18 \times$ total internet traffic in 2000
- UK National Health Service plans to sequence genome of 750.000 cancer patients in the next ten years
- Google purchased DeepMind (after 1 year of operation) for 450M GBP
- Number of job postings for data scientists increased globally by 15.000% between 2011 and 2012

The problem

- Vast amounts of quantitative data arising from every aspect of life
- Advanced informatics tools necessary just to handle the data
- Widespread belief that data is valuable, yet worthless without analytic tools
- Converting data to knowledge is the challenge

A memorable quote

- *If you ignore philosophy, it comes back and bites your bottom* (Dr R. Shillcock, Informatics, Edinburgh)
- What is a model? Discuss for 5 minutes and provide 3 examples

My own answer

- A model is a hypothesis that certain features of a system of interest are well replicated in another, simpler system.
- A *mathematical model* is a model where the simpler system consists of a set of mathematical relations between objects (equations, inequalities, etc).
- A *stochastic model* is a mathematical model where the objects are probability distributions.
- All modelling usually starts by defining a *family* of models indexed by some parameters, which are tweaked to reflect how well the feature of interest is captured.
- Machine learning deals with algorithms for automatic selection of a model from observations of the system.

My own answer

- A model is a hypothesis that certain features of a system of interest are well replicated in another, simpler system.
- A *mathematical model* is a model where the simpler system consists of a set of mathematical relations between objects (equations, inequalities, etc).
- A *stochastic model* is a mathematical model where the objects are probability distributions.
- All modelling usually starts by defining a *family* of models indexed by some parameters, which are tweaked to reflect how well the feature of interest is captured.
- Machine learning deals with algorithms for automatic selection of a model from observations of the system.

My own answer

- A model is a hypothesis that certain features of a system of interest are well replicated in another, simpler system.
- A *mathematical model* is a model where the simpler system consists of a set of mathematical relations between objects (equations, inequalities, etc).
- A *stochastic model* is a mathematical model where the objects are probability distributions.
- All modelling usually starts by defining a *family* of models indexed by some parameters, which are tweaked to reflect how well the feature of interest is captured.
- Machine learning deals with algorithms for automatic selection of a model from observations of the system.

My own answer

- A model is a hypothesis that certain features of a system of interest are well replicated in another, simpler system.
- A *mathematical model* is a model where the simpler system consists of a set of mathematical relations between objects (equations, inequalities, etc).
- A *stochastic model* is a mathematical model where the objects are probability distributions.
- All modelling usually starts by defining a *family* of models indexed by some parameters, which are tweaked to reflect how well the feature of interest is captured.
- Machine learning deals with algorithms for automatic selection of a model from observations of the system.

My own answer

- A model is a hypothesis that certain features of a system of interest are well replicated in another, simpler system.
- A *mathematical model* is a model where the simpler system consists of a set of mathematical relations between objects (equations, inequalities, etc).
- A *stochastic model* is a mathematical model where the objects are probability distributions.
- All modelling usually starts by defining a *family* of models indexed by some parameters, which are tweaked to reflect how well the feature of interest is captured.
- Machine learning deals with algorithms for automatic selection of a model from observations of the system.

Course content

References to Barber 2010 book.

- Lecture 1: Statistical basics. Probability refresher, probability distributions, entropy and KL divergence (Ch 1, Ch 8.2, 8.3). Multivariate Gaussian (8.4). Estimators and maximum likelihood (8.6 and 8.7.3). Supervised and unsupervised learning (13.1)
- Lecture 2: Linear models. Regression with additive noise and logistic regression (probabilistic perspective): maximum likelihood and least squares (18.1 and 17.4.1). Duality and kernels (17.3).
- Lecture 3: Bayesian regression models and Gaussian Processes. Bayesian models and hyperparameters (18.1.1, 18.1.2). Gaussian Process regression (19.1-19.4).
- Lecture 4: Active learning and Bayesian optimisation. Active

Course content cont'd

- Lecture 5: Latent variables and mixture models. Latent variables and the EM algorithm (11.1 and 11.2.1). Gaussian mixture models and mixture of experts (20.3, 20.4).
- Lecture 6: Graphical models. Belief networks and Markov networks (3.3 and 4.2). Factor graphs (4.4).
- Lecture 7: Exact inference in trees. Message passing and belief propagation (5.1 and 28.7.1).
- Lecture 8: Approximate inference in graphical models. Variational inference: Gaussian and mean field approximations (28.3, 28.4). Sampling methods and Gibbs sampling (27.4 and 27.3).
- Lab 2: Bayesian Gaussian Mixture Models

Definitions

- Random variables: results of non exactly reproducible experiments
- Either intrinsically random (e.g. quantum mechanics) or the system is incompletely known, cannot be controlled precisely
- The probability p_i of an experiment taking a certain value i is the frequency with which that value is taken in the limit of infinite experimental trials
- Alternatively, we can take probability to be our belief that a certain value will be taken

More definitions

- Let x be a random variable, the set of possible values of x is the *sample space* Ω
- Let x and y be two random variables, $p(x = i, y = j)$ is the *joint probability* of x taking value i and y taking value j (with i and j in the respective sample spaces. Often just written $p(x, y)$ to indicate the function (as opposed to its evaluation over the outcomes i and j)
- $p(x|y)$ is the conditional probability, i.e. the probability of x if you know y has a certain value

Rules

- *Normalisation*: the sum of the probabilities of all possible experimental outcomes must be 1, $\sum_{x \in \Omega} p(x) = 1$
- *Sum rule*: the marginal probability $p(x)$ is given by summing the joint $p(x, y)$ over all possible values of y,

$$p(x) = \sum_{y \in \Omega} p(x, y)$$

- *Product rule*: the joint is the product of the conditional and the marginal, $p(x, y) = p(x|y)p(y)$
- *Bayes rule*: the posterior is the ratio of the joint and the marginal

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Distributions and expectations

- A probability distribution is a rule associating a number $0 \leq p(x) \leq 1$ to each state $x \in \Omega$, such that $\sum_{x \in \Omega} p(x) = 1$
- For finite state space can be given by a table, in general is given by a functional form
- Probability distributions (over numerical objects) are useful to compute expectations of functions

$$\langle f \rangle = \sum_{x \in \Omega} f(x) p(x)$$

- Important expectations are the *mean* $\langle x \rangle$ and *variance* $var(x) = \langle (x - \langle x \rangle)^2 \rangle$. For more variables, also the *covariance* $cov(x, y) = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$ or its scaled relative the *correlation* $corr(x, y) = cov(x, y) / \sqrt{var(x) var(y)}$

Computing expectations

- If you know analytically the probability distribution and can compute the sums (integrals), no problem
- If you know the distribution but cannot compute the sums (integrals), enter the magical realm of approximate inference (fun but out of scope)
- If you know nothing but have N_S samples, then use a sample approximation
- Approximate the probability of an outcome with the *frequency* in the sample

$$\langle f(x) \rangle \simeq \sum_x \frac{n_x}{N_S} f(x) = \frac{1}{N_S} \sum_{i=1}^{N_S} f(x_i)$$

(prove the last equality)

Independence

- Two random variables x and y are *independent* if their joint probability factorises in terms of marginals

$$p(x, y) = p(x)p(y)$$

- Using the product rule, this is equivalent to the conditional being equal to the marginal

$$p(x, y) = p(x)p(y) \Leftrightarrow p(x|y) = p(x)$$

- Exercise: if two variables are independent, then their correlation is zero. **NOT TRUE** viceversa (no correlation does not imply independence)

Continuous states

- If the state space Ω is continuous some of the previous definitions must be modified
- The general case is mathematically difficult; we restrict ourselves to $\Omega = \mathbb{R}^n$ and to distributions which admit a *density*, a function

$$p : \Omega \to \mathbb{R} \quad \text{s.t.} \quad p(x) \geq 0 \forall x \quad \text{and} \quad \int_{\Omega} p(x)dx = 1$$

- It can be shown that the rules of probability distributions hold also for probability densities
- Notice that $p(x)$ is NOT the probability of the random variable being in state x (that is always zero for bounded densities); probabilities are only defined as integrals over subsets of Ω

Entropy and divergence

- Probability theory is the basis of information theory (interesting, but not the topic of this course).
- An important quantity is the *entropy* of a distribution

$$H[p] = -\sum_i p_i \log_2 p_i$$

- Entropy measures the level of disorder of a distribution; for discrete distributions, it is always ≥ 0 and 0 only for deterministic distirbutions
- The *relative entropy* or *Kullback-Leibler (KL) divergence* between two distributions is

$$KL[q\|p] = \sum_i q_i \log \frac{q_i}{p_i}$$

- Fact: KL is convex and ≥ 0

Basic distributions

- Discrete distribution: a random variable can take N distinct values with probability $p_i, \quad i = 1, \ldots, N$. Formally

$$p(x = i) = \prod_j p_j^{\delta_{ij}}$$

  δ_{ij} is the Kronecker delta and the p_is are parameters.
- Poisson distribution: a distribution over non-negative integers

$$p\left(n|\mu\right) = \frac{\mu^n}{n!} \exp[-\mu]$$

  The parameter μ is often called the *rate* of the distribution.
- The Poisson distribution is often used for *rare events*, e.g. decaying of particles of *binding of DNA fragments to a probe* (more later!)
- Exercise: compute mean and variance of a Poisson distribution

Basic distributions

- Multivariate normal: distribution over vectors $\mathbf{x}$, density

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

  $\boldsymbol{\mu}$ is the mean and Σ is the covariance matrix. Often useful to parametrise it in terms of the *precision matrix* Σ^{-1}. How many parameters does a multivariate normal have?

- Gamma distribution: distribution over positive real numbers, density

$$p(x|k, \theta) = \frac{x^{k-1}\exp(-x/\theta)}{\theta^k \Gamma(\theta)}$$

  with shape parameter k and scale parameter θ

Interesting exercise

This exercise illustrates the pitfalls of working in high dimensions.

1. *Curse of dimensionality*: Suppose you want to explore uniformly a region by gridding it. How many grid points do you need?
2. *Even worse*: Suppose you sample from a spherical Gaussian distribution. Where do the points lie as the dimensions increase?

Mixtures: how to build more distributions

- More general distributions can be built via mixtures: e.g.

$$p(x|\mu_{1...,n}, \sigma^2_{1,...,n}) = \sum_i \pi_i \mathcal{N}(\mu_i, \sigma_i^2)$$

  where the *mixing coefficients* π_i are discretely distributed
- You can interpret this as a two stage hierarchical process: choose one component out of a discrete distribution, then choose the distribution for that component
- **IMPORTANT CONCEPT**: this is an example of *latent variable model*, with a latent class variable and an observed continuous value. The mixture is the marginal distribution for the observations
- The probability of the latent variables given the observations can be obtained using Bayes' theorem: see next week

Continuous mixtures: some cool distributions

- No need for the mixing distribution (latent variable) to be discrete
- Suppose you are interested in the means of normally distributed samples (possibly with different variances/ precisions)
- Marginalising the precision in a Gaussian using a Gamma mixing distribution yields a *Student t-distribution*
- Suppose you have multiple rare event processes happening with slightly different rates
- Marginalising the rate in a Poisson distribution using a Gamma mixing distribution yields a *negative binomial* distribution

Parameters?

- Many distributions are written as conditional probabilities *given* the parameters
- Often the values of the parameters are not known
- Given independent and identically distributed (i.i.d.) observations, we can estimate them; e.g., we pick θ by maximum likelihood

$$\hat{\theta} = \text{argmax}_\theta \left[\prod_i p(x_i|\theta) \right]$$

- Alternatively, you could have a prior over the parameters $p(\theta)$ and take the *maximum a posteriori (MAP)* estimate

$$\hat{\theta}_{MAP} = \text{argmax}_\theta \left[p(\theta) \prod_i p(x_i|\theta) \right]$$

Justification for maximum likelihood

- Given a data set $\{x_i\}, \quad i = 1, \ldots, N$, let the empirical distribution be

$$p_{emp}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(x_i)$$

with $\mathbb{I}$ the indicator function of a set

- To find a suitable distribution q to model the data, one may wish to minimize the Kullback-Leibler divergence

$$KL[p_{emp} \| q] = H[p_{emp}] - \langle \log q(x) \rangle_{p_{emp}} = -\frac{1}{N} \sum \log q(x_i)$$

- *Maximum likelihood is equivalent to minimizing a KL divergence with the empirical distirbution*

Exercise: fitting a discrete distribution

- We have independent observations $x_1, \ldots, x_N$ each taking one of D possible values, giving a likelihood

$$\mathcal{L} = \prod_{i=1}^{N} p\left(x_i|\mathbf{p}\right)$$

- Compute the Maximum Likelihood estimate of $\mathbf{p}$. What is the intuitive meaning of the result? What happens if one of the D values is not represented in your sample?

Exercise II: fitting a Gaussian distribution

- We have independent, real valued observations $x_1, \ldots, x_N$
- Find the parameters of the optimal Gaussian fit by maximum likelihood

Bayesian estimation

- The Bayesian approach quantifies uncertainty at every step
- The parameters are treated as additional random variables with their own *prior* distribution $p(\theta)$
- The observation likelihood is combined with the prior to obtain a *posterior* distribution via Bayes' theorem

$$p(\theta|x_{\mathcal{I}}) = \frac{p(x_{\mathcal{I}}|\theta)p(\theta)}{p(x_{\mathcal{I}})}$$

where $\mathcal{I}$ is the set indexing the observations

- The distribution of the observable x (*predictive* distribution) is obtained as

$$p(x|x_{\mathcal{I}}) = \int d\theta p(x|\theta)p(\theta|x_{\mathcal{I}})$$

Exercise: Bayesian fitting of Gaussians

- Let data $x_i \quad i = 1, \ldots, N$ be distributed according to a Gaussian with mean μ and variance σ^2
- Let the prior distribution over the mean μ be a Gaussian with mean m and variance v^2
- Compute the posterior and predictive distribution

Estimators

- A procedure to calculate an expectation is called an *estimator*
- e.g., fitting a Gaussian to data by maximum likelihood provides the M.L. estimator for mean and variance, or Bayesian posterior mean
- An estimator will be a noisy estimate of the true value, due to finite sample effects
- An estimator $\hat{f}$ is *unbiased* if its expectation (under the joint distribution of the data set) coincides with the true value
- Exercise: show that the ML estimator of variance is biased

Learning as fitting distributions

- The world-view of this course is that a model consists of a set of random variables and probabilistic relationships describing their interactions
- *Learning* refers to computing conditional distributions w.r.t. some observations of subsets of the model
- *Predictions* are then carried out using the Bayesian predictive distribution

Supervised and unsupervised learning

- Slightly meaningless terms still heavily used
- Focus on models involving more than one random variable
- In particular, supervised learning applies when data is in the form of *input-output pairs*
- Supervised learning aims at learning the (probabilistic) functional relationship between the output and the input
- Unsupervised learning refers to purely learn the structure of the probability distribution underlying the data

Generative and discriminative models

- Supervised learning can have two flavours
- Two different types of question can be asked:
  - what is the joint probability of input/ output pairs?
  - given a new input, what will be the output?
- The first question requires a model of the population structure of the inputs, and of the conditional probability of the output given the input $\longrightarrow$ **generative modelling**
- The second question is more parsimonious but less explanatory $\longrightarrow$ **discriminative learning**
- Notice that the difference between generative supervised learning and unsupervised learning is moot