LECTURE 1
Course preliminaries

The term machine learning goes back to Arthur Samuels and his computer checker playing algoriths. In 1959 Samuels described machine learning as: ”Field of study that gives computers the ability to learn without being explicitly programmed.”

Machine learning is considered a subfield of artificial intelligence and the idea of a learning machine is given in ”Computing Machinery and Intelligence,” by Alan Turing in 1950 in Mind: A Quarterly Review of Psychology and Philosophy. The question posed in the fist sentence of this paper was “Can machines think ?”.

For this class by ML we are going to consider algorithms and probabilistic methods to ”learn from data.” The material is at the interface of statistics and computer science and one caricature of ML is computer scientists doing statistics. ML is often also associated with the term ”big data” which is often meant to be statistical analysis with very large data sets, here the computational challenge is as serious as the inference problem.

Broadly speaking the methods we will discuss can be placed into two categories:
proceduralists: This will cover both frequentist statistics, as well as algorithmic approaches to ML. This approach is based upon coming up with good procedures to apply to data. What is meant by good is some long run probability of the procedure, for example the long run probability of errors made in classification is small.
Bayesian: A coherent axiomatic approach to inference based on inference of the posterior probability of parameters or models given data. Bayesian inference may not be feasible or practical in certain situations.

1.1. Review

We'll start with a basic review of statistics. We will examine a statistical question using both Bayesian and frequentist analysis. The following formalism will be quantified in both models

$$\begin{aligned}
\mathrm{P}(M \mid D) &= \frac{\mathrm{P}(D \mid M)\mathrm{P}(M)}{\mathrm{P}(D)} \\
&\propto \mathrm{P}(D \mid M)\mathrm{P}(M),
\end{aligned}$$

where $\mathrm{P}(M \mid D)$ is evidence for model M given data D, $\mathrm{P}(D \mid M)$ is evidence for D given model M, $\mathrm{P}(M)$ is the probability of model M, and $\mathrm{P}(D)$ the probability of data, The standard statistical terms for these objects are

$$\begin{aligned}
\mathrm{P}(D \mid M) &\equiv \mathrm{Lik}(D; M), \quad \text{Likelihood of data given model } M \\
\mathrm{P}(M \mid D) &\equiv \mathrm{Post}(D; M), \quad \text{Posterior evidence of model } M \text{ given data} \\
\mathrm{P}(M) &\equiv \pi(M), \quad \text{prior probability (before seeing data) for model } M.
\end{aligned}$$

Example 1: Motif estimation

We consider a random variable X that is drawn from a alphabet of $k = 4$ letters $\{A, C, T, G\}$ where we represent $A \equiv 1$, $C \equiv 2$, $T \equiv 3$, and $G \equiv 4$. We set the probability distribution on X as the following multinomial distribution, note we are modeling a draw of four letters

$$\begin{aligned} \mathrm{P}(n_1, n_2, n_3, n_4 \mid p_1, p_2, p_3, p_4) &\equiv \mathrm{Multi}(p_1, p_2, p_3, p_4) \\ &\propto \prod_{j=1}^{4} p_j^{n_j}, \quad \sum_{j=1}^{4} p_j = 1, p_j \geq 0 \ \forall j = 1, ..., 4, \end{aligned}$$

where p_i is the probability of observing the i-th letter ($\{A, C, T, G\}$ in the alphabet and n_i states how many times the i-th letter is observed (either 1 or 0). The above is an example of the multinomial distribution.

The random variable X is a string in a sequence and we can think of the random string $Z = (X_1, ..., X_m)$ as a string of length m with each X_i drawn iid from a distribution. This is an example of a string, let us call these strings motifs. The data consists of a series of n strings, $D = \{Z_1, ..., Z_n\}$ with each string Z_i drawn iid (independently and identically distributed).

We first state the likelihood of observing the data D

$$\begin{aligned} \mathrm{P}(D \mid M) &= \mathrm{Lik}(D \mid p_1, ..., p_4) \\ \mathrm{Lik}(D \mid p_1, ..., p_4) &\propto \prod_{i=1}^{n} \left[\prod_{\ell=1}^{m} \left(\prod_{j=1}^{k} p_j^{n_{i\ell j}} \right) \right] \\ &\propto \prod_{\ell=1}^{m} \left[\prod_{i=1}^{n} \left(\prod_{j=1}^{k} p_j^{n_{i\ell j}} \right) \right] \\ &\propto \prod_{\ell=1}^{m} \left[\prod_{j=1}^{k} p_j^{\tilde{n}_{\ell j}} \right], \end{aligned}$$

where $n_{i\ell j}$ is the number of observations of letter j at position ℓ in observation i (again this is 0 or 1) and $\tilde{n}_{\ell j} = \sum_i n_{i\ell j}$ is the number of times in the n sequences that letter j is observed at position ℓ.

A classical method for estimating $p_1, .., p_k$ is the maximum likelihood formulation

$$\begin{aligned} \{\hat{p}_1, ..., \hat{p}_k\} &= \arg \max_{p_1, \cdots, p_k} \left[\mathrm{Lik}(D \mid p_1, ..., p_k) \right], \\ &\qquad \text{subject to} \sum_{j=1}^{k} p_j = 1, p_j \geq 0 \ \forall j = 1, ..., k. \end{aligned}$$

To understand how to do the above optimization learn about the method of lagrange multipliers. This is a very reasonable approach but it has one problem, how does one estimate the uncertainty in the estimate of $\{\hat{p}_1, ..., \hat{p}_k\}$?

We can formally model the uncertainty using Bayes rule

$$\mathrm{P}(M \mid D) \propto \mathrm{P}(D \mid M)\mathrm{P}(M),$$

if we can put a probability distribution on the model space, in this case $(p_1, ..., p_k)$. The space of all points $\mathbf{p} = (p_1, ..., p_k)$ such that $\sum_j p_k = 1$ and $p_k \geq 0$ for all $j = 1, ..., k$ is called the simplex. We now state a classical distribution on the simplex called the Dirichlet distribution

$$\begin{aligned} f(p_1, ..., p_k \mid \alpha_1, ..., \alpha_k) &\equiv \mathrm{Dir}(\alpha_1, ..., \alpha_k) \\ &\propto \prod_{j=1}^{k} p_j^{\alpha_j - 1}, \quad \alpha_j \geq 0 \, \forall j, \, \alpha_j \in \mathbb{N}, \end{aligned}$$

where $\mathbb{N}$ are the natural numbers, it is natural to think of the $\{\alpha_1, ..., \alpha_k\}$ parameters as counts. We can use the Dirichlet distribution as a prior $\pi(M)$ with the uniform prior being $\mathrm{Dir}(\alpha_1 = 1, ..., \alpha_k = 1)$. We now state the posterior

$$\begin{aligned} \mathrm{P}(M \mid D) &\propto \mathrm{Lik}(D \mid p_1, ..., p_4) \times \pi(p_1, ..., p_4) \\ &\propto \prod_{i=1}^{n} \left[\prod_{\ell=1}^{m} \left(\prod_{j=1}^{k} p_j^{n_{i\ell j}} \right) \right] \times \prod_{j=1}^{k} p_j^{\alpha_j - 1} \\ &\propto \prod_{\ell=1}^{m} \left[\prod_{i=1}^{n} \left(\prod_{j=1}^{k} p_j^{n_{i\ell j}} \right) \right] \times \prod_{j=1}^{k} p_j^{\alpha_j - 1} \\ &\propto \prod_{\ell=1}^{m} \left[\prod_{j=1}^{k} p_j^{\tilde{n}_{\ell j}} \right] \times \prod_{j=1}^{k} p_j^{\alpha_j - 1} \\ &\propto \prod_{\ell=1}^{m} \left[\prod_{j=1}^{k} p_j^{\tilde{n}_{\ell j}} \right] \times \prod_{j=1}^{k} p_j^{\alpha_j - 1} \\ &\propto \left[\prod_{j=1}^{k} p_j^{\breve{n}_j} \right] \times \prod_{j=1}^{k} p_j^{\alpha_j - 1} \\ &\propto \left[\prod_{j=1}^{k} p_j^{\breve{n}_j + \alpha_j - 1} \right] \\ &= \mathrm{Dir}(\breve{n}_1 + \alpha_1, ..., \breve{n}_k + \alpha_k), \end{aligned}$$

where $\breve{n}_j = \sum_{i\ell} n_{i\ell j}$. The strength of this estimation procedure is that we end up with not just a point estimate $\{\hat{p}_1, ..., \hat{p}_k\}$ as we did in the MLE approach but we end up with a posterior distribution. We can use the highest probability value for $(p_1, ..., p_k)$ as an estimate or the mean of the posterior distribution. The reason why this example worked so easily is that the multinomial and Dirichlet distributions are conjugate. By this we mean that

$$\mathrm{Multi}(p_1, ..., p_k) \times \mathrm{Dir}(\alpha_1, ..., \alpha_k) = \mathrm{Dir}(p_1 + \alpha_1, ..., p_k + \alpha_k).$$