ISML-II: Machine Learning **Spring 2014**

Lecture 24- Bayesian Learning

Lecturer: Lorenzo Rosasco

The point of view taken thus far is to try whenever possible to estimate only the quantity of interest rather than the underlying probability distribution, and iny case trying to make as little assumptions as possible on the latter. This perspective is in contrast with Bayesian approaches where probability distribution are the central objects of interest. We next provide some basic idea for this class of methods.

24.1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is one of the most classical estimation principles. The basic idea of maximum likelihood estimation is to derive the assume a parametric model for the unknown the distribution and then estimate the parameter which mostly likely generated the data.

24.2 Maximum Likelihood for Density Estimation

Consider the problem of estimating a density function p on $\mathbb{R}$, from independent samples $x_1, \dots, x_n$. In contrast to nonparametric approaches, here we make explicit parametric assumptions on the uknown distribution. In particular, we begin assuming p to be a standard Gaussian, so that the problem reduces to the estimation of the mean. In the following we denote by p_μ the density to explicitly indicate the dependence on the unknown mean.

Since the data are independent, their joint distribution, called *likelihood*, is the product of the individual distributions

$$\prod_{i=1}^n p_\mu(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-|x_i-\mu|^2} = e^{-\frac{1}{\sqrt{2\pi}} \sum_{i=1}^n |x_i-\mu|^2}$$

where we used the Gaussian assumption in the last equalities. The log-likelihood is simply the logarithm of the above quantity

$$\log \prod_{i=1}^n p_\mu(x_i) \propto - \sum_{i=1}^n |x_i - \mu|^2.$$

The idea is then to choose the value of the mean that maximize the likelihood (and hence the joint probability of observing the given data). We add several remarks.

- In the above example, it is straightforward to see that the best MLE of the mean is

$$\mu_n = \frac{1}{n} \sum_{i=1}^n x_i.$$

- The above example can be extended to consider more unknown parameters, for example both the mean and the variance. In this case, the likelihood is a multivariate function (a function of many variables).
- The above example can also be easily generalized to multiple dimensions considering a multivariate Gaussian
$$N(\mu, I) = Ce^{-\|x_i-\mu\|^2}$$
where C is a normalizing constant. In this case, the log-likelihood is proportional to
$$-\sum_{i=1}^{n} \|x_i - \mu\|^2 .$$
- Finally, the above example can generalized to parametric assumptions p_θ other than Gaussian, where θ is a vector of unknown parameters. In this case the likelihood is a function of θ.

24.3 Maximum Likelihood for Linear Regression

The above approach can be applied to linear regression. Consider the regression model

$$y_i = w^\top x_i + \epsilon_i, \qquad i = 1, \dots, n,$$

and assume for all $i = 1, \dots, n$, that ϵ_i are independent samples of a Gaussian $N(0, \sigma^2)$. Here the variance $\sigma > 0$ can be seen as a noise level. The goal is to estimate the unknown coefficient vector w_*, given $(x_1 y_1), \dots, (x_n, y_n)$.

Following a MLE approach, for all $i = 1, \dots, n$, we are going to consider the conditional distribution

$$p(y_i, |x_i; w)$$

given by the Gaussian $N(w^\top x_i, \sigma^2) = Ce^{-\frac{1}{2\sigma^2}|w^\top x_i - y_i|^2}$, where C is a normalizing constant. Then, because of independence,

$$p(y_1, \dots, y_n | w^*, x_1, \dots, x_n; w) = -\prod_{i=1}^{n} Ce^{-\frac{|w^\top x_i - y_i|^2}{2\sigma^2}} = C' e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}|w^\top x_i - y_i|^2}$$

where C' is a normalizing constant. Then the likelihood is proportional to

$$-\frac{1}{\sigma^2}\sum_{i=1}^{n}|w^\top x_i - y_i|^2$$

and we see that maximizing the likelihood reduces to the least squares.

In other words, least squares can be derived from a MLE approach considering a linear regression model under a Gaussian noise model.

24.4 Prior and Posterior

The basic idea of Bayesian estimation is that the parameter of interest follow a prior distribution reflecting our prior knowledge/beliefs on the problem. The idea is then to derive the so called posterior distribution, obtained conditioning the prior to the observed data. The basic tool is the Bayes rule, which, given two random variables U, V, in its simplest form, is expressed by the equality

$$P(U|V) = \frac{P(V|U)P(U)}{P(V)}.$$

24.4.1 Posterior for Linear Regression

We illustrate the above idea in the case of linear regression. In this case, a classic prior is expressed by the assumption that the unknown coefficient vector w is distributed according to a standard multivariate Gaussian $N(0, I)$.

The idea is to apply the Bayes rule to the data and the coefficient vector w seen as random quantities. More precisely we have

$$\begin{aligned} P(w \mid y_1, \dots, y_n; x_1, \dots, x_n) &= \tfrac{P(y_1,\dots,y_n;x_1,\dots,x_n \mid w)P(w)}{P(y_1,\dots,y_n;x_1,\dots,x_n)} \\ &= \tfrac{P(y_1,\dots,y_n \mid x_1,\dots,x_n;w)P(w)}{P(y_1,\dots,y_n|x_1,\dots,x_n)}, \end{aligned}$$

or using a more compact vector notation

$$P(w \mid X_n, Y_n) = \frac{P(Y_n \mid X_n, w)P(W)}{P(Y_n|X_n,)},$$

where X_n denotes the n by D inputs matrix and Y_n the n dimensional vector of corresponding outputs. Several comments can be made.

- The above expression can be interpreted as determining how our prior beliefs, $P(w)$, are updated once we observe the data, $P(w|y_1, \dots, y_n; x_1, \dots, x_n)$. The latter quantity is called posterior probability.
- From the Bayes rule we have that the posterior depends on the product of likelihood and the prior divided by the, so called, marginal likelihood.
- The marginal likelihood is only a normalizing factor since it does not depend on w.

Given the Gaussian assumptions on the noise and coefficient vector, the posterior expression can be given explicitly, as we discuss next.

From the above discussion we have that

$$P(w \mid X_n, Y_n) \propto P(Y_n \mid X_n, w)P(W)$$

The first term in the right end side can be identified with the likelihood

$$P(Y_n \mid X_n, w) = Ce^{-\frac{\sum_{i=1}^n \left\|w^T x_i - y_i\right\|^2}{2\sigma^2}},$$

whereas the second term is the prior

$$P(w) = C' e^{-\|w\|^2},$$

for suitable normalizing constants C, C'. This means that the posterior is given up to a normalizing constant by

$$e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n |w^T x_i - y_i|^2} e^{-\|w\|^2} = e^{-\left(\frac{1}{2\sigma^2}\sum_{i=1}^n |w^T x_i - y_i|^2 + \|w\|^2\right)}.$$

In words, the posterior is given by the product of two Gaussian distributions and a possible parameter estimate is given by computing its maximum (which is also equal to its mean). This parameter estimate is called Maximum A Posteriori (MAP) and is given by the solution of the problem

$$\max_{w \in \mathbb{R}^D} e^{-\left(\frac{1}{2\sigma^2}\sum_{i=1}^n |w^T x_i - y_i|^2 + \|w\|^2\right)}.$$

As in the case of the MLE for linear regression with Gaussian noise model, also the MAP estimates corresponding to the Gaussian prior provides an interesting connection. Indeed the above problem is equivalent to

$$\min_{w \in \mathbb{R}^D} \frac{1}{2\sigma^2}\sum_{i=1}^n |w^T x_i - y_i|^2 + \|w\|^2.$$

In other words ridge regression can be derived as the MAP estimate for linear regression under a Gaussian prior and Gaussian noise model. In this case the regularization parameter is determined by the noise level.

Beyond the maximum (or mean) parameter, the interest of having an explicit expression for the posterior is that other moments can be computed, for example the variance of the parameter estimate. We won't develop further these computations, but only point out how the possibility of estimating this higher order information is a hall-mark of Bayesian approaches, where it is referred to as uncertainty quantification.

24.5 Beyond Linear Regression

The above ideas can be extended in a number of different ways.

Non-linear parametric regression. It is straightforward to see that the above reasoning and calculations extend to the case where we consider a regression model

$$y_i = \sum_{j=1}^p w_*^j \phi_j(x_i) + \epsilon_i, \qquad i = 1, \dots, n,$$

for some set of p features $\phi_j : \mathbb{R}^D \to \mathbb{R}$, $j = 1, \dots, p$. Indeed, all the above estimates hold simply replacing x_i by $\tilde{x}_i = (\phi_1(x_i), \dots, \phi_p(x_i))$ for all $i = 1, \dots, n$.

Non-linear nonparametric regression:Gaussian Processes. It is also possible to consider the extension of the above reasoning to an infinite set of features essentially deriving a parallel of kernel methods from a Bayesian perspective. This derivation, based on Gaussian processes, is outside the scope of our presentation.

From regression to classification. Finally the above reasoning can also be considered in the case of classification. Following the Bayesian ansatz a suitable probabilistic model needs be specified in this case since the regression model with Gaussian noise is meaningless in this case. While this is indeed possible, in general the corresponding computations are more involved. Also this topic left out from this presentation.