Statistical Machine Learning

UoC Stats 37700, Winter quarter

Lecture 11: Bayesian approaches, Gaussian processes

The Bayesian way

- The general Bayes principle is to assume a prior distribution on some space of parameters $\theta \in \Theta$ that determine the data generating distribution:

$$\theta \sim \pi \rightarrow \mathbb{P}_\theta \in \mathcal{P} \rightarrow (Z_1, \ldots, Z_n) \sim \mathbb{P}_\theta^{\otimes n} .$$

- Given the observed data $(Z_1, \ldots, Z_n)$, the main object of interest is the posterior distribution on the parameters:

$$\begin{aligned} p(\theta | Z_1, \ldots, Z_n) &= \frac{p(Z_1, \ldots, Z_n | \theta) p(\theta)}{p(Z_1, \ldots, Z_n)} \\ &= \pi(\theta) . \prod_{i=1}^{n} \mathbb{P}_\theta [Z_i] . \mathbb{E}_{\theta \sim \pi} \left[\prod_{i=1}^{n} \mathbb{P}_\theta [Z_i] \right]^{-1} \end{aligned}$$

Latent variables

- Often the parameter θ parametrizes a "latent variable"

$$V = f_\theta(Z) \text{ and } Z|V \sim \mathbb{Q}_V$$

(where $\mathbb{Q}_t$ is a fixed family).

- For regression $Z = (Z, Y)$, the usual model

$$Y_i = f_\theta(X_i) + \varepsilon_i \,,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

- For classification, we can use logistic modelling,

$$Y_i = \mathit{Bernoulli}(\text{sigmoid}(f_\theta(X_i))); \quad \text{sigmoid}(x) = \frac{1}{1+\exp(-x)} \,.$$

Using a basis of fixed functions

- Consider the regression or classification case where the parametrization is given by

$$f_\theta(x) = \sum_{i=1}^{K} \theta_i h_i(x)\,,$$

  where (h_i) form a basis of fixed in advance functions.
- Assume the prior over the parameters is given by a $\mathcal{N}(0, \gamma^2)$ distribution.

Maximum A Posteriori

- A common way to perform Bayesian inference is to find the maximum of the a posteriori (MAP) distribution.
- For regression, in the previous framework this results in finding

$$\min_{\theta} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - f_\theta(X_i))^2 + \frac{1}{2\gamma^2} \|\theta\|^2 ,$$

  (where terms independant of θ have been dropped)
- For binary classification,

$$\min_{\theta} \sum_{i=1}^{n} \log(1 + \exp(-Y_i f_\theta(X_i))) + \frac{1}{2\gamma^2} \|\theta\|^2 .$$

- Thus MAP estimation is equivalent to regularized empirical loss minimization.

The real Bayes way: predictive distribution

- The MAP is however not really satisfying from a Bayesian standpoint (for example it is not invariant wrt. reparametrization, or change of the reference measure).
- The *predictive* Bayesian framework uses the current information to determine the amount of uncertainty about a new sample point Z^*:

$$p(Z^*|Z_1,\ldots,Z_n) = \int_\theta \mathbb{P}_\theta(Z^*)p(\theta|Z_1,\ldots,Z_n)\,.$$

- In the case of classification/regression, the observation X^* is known and we want to infer Y^*, so that (generally) the above is rewritten with everything conditional on the design $(X_1,\ldots,X_n,X^*)$ (in other words there is no modelling of the design).
- The advantage of the predictive distribution is that it gives an idea of the uncertainty about the prediction (via the predictive distribution). Of course, this notion of uncertainty strongly depends on the prior that has been picked in the first place.

Computable example

- In the regression case, the Bayesian model is

$$Y = H\Theta + \varepsilon\,,$$

where $H_{ij} = h_j(X_i)\,,\ \Theta \sim \mathcal{N}(0, \gamma^2 I_K)\,,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)\,.$

- The posterior ir proportional to

$$p(\Theta|(X_i, Y_i)_{1\leq i\leq n}) \propto \exp\left(-\frac{1}{2\sigma^2}(\Theta - M_\nu^{-1}HY)^t M_\nu(\Theta - M_\nu^{-1}HY)\right)\,,$$

where $M_\nu = H^t H + \nu^2 I_K\,,\ \nu^2 = \sigma^2/\gamma^2\,.$

- Hence

$$\Theta|((X_i, Y_i)_{1\leq i\leq n}) \sim \mathcal{N}(M_\nu^{-1}HY\,, \gamma^2 M_\nu^{-1})$$

- The predictive distribution of Y^* given a new observation X^* is given by left multiplication by the row vector $h^* = (h_1(X^*), \ldots, h_K(X^*))\,.$

- Alternative derivation: $\begin{pmatrix} Y \\ \Theta \end{pmatrix}$ is jointly centered Gaussian with covariance matrix

$$\gamma^2 \begin{pmatrix} K_\nu & H \\ H^t & I_K \end{pmatrix},$$

where $K_\nu = HH^t + \nu^2 I_n$.

- using the conditioning formula for Gaussians we obtain

$$\Theta | (X_i, Y_i)_{1 \leq i \leq n} \sim \mathcal{N}(H^t K_\nu^{-1} Y, \gamma^2 (I_K - H^t K_\nu^{-1} H)),$$

- (Compare with previous expression)

“Hyperparameters”

- In the previous model σ^2, γ^2 are arbitrary parameters.
- We can put a prior on those parameters too: hierarchical Bayes.
- Note that this is just a convienient way to specify a more general prior over Θ, Y (which is, in particular, non-Gaussian)

How can we compute the predictive distribution?

We can't in general. However, there are a number of approximation methods available. In fact, much of the area of research concerning Bayesian approaches is dedicated to those.

- perform a MAP on Hyperparameters (marginal likelihood or evidence maximization), then fix them and perform Bayesian prediction as if they were fixed.
- Approximate distribution of hyperparameters by a Gaussian centered at the MAP value (Laplace approximation).
- Various forms of Monte-Carlo integration.
- Approximate the posterior via a more tractable family of distributions having a specific form, for example: the different parameters are independent in the posterior (variational Bayes, unfortunately also sometimes called ensemble learning).

Marginal likelihood/evidence maximization

- Marginal likelihood maximization or “type II MAP” selects hyperparameters (here σ, γ) by looking at

$$p((Z_i)_{1\leq i\leq n}|\sigma,\gamma) = \int \mathbb{P}_\theta\left[(Z_i)\right] p(\theta|\sigma,\gamma) d\theta$$

- The Bayesian credo is that this will result in a tradeoff between model complexity and available information from the data. The intuitive view is that simple models will give a high likelihood to only a limited set of “typical” samples, while more complex models “spread” the likelihood over more samples (thereby also diluting it, hence the tradeoff).
- This view, the Bayesian’s Occam’s razor, works quite well in practice but only if the number of parameters to choose this way is limited, otherwise we are again at risk of overfitting.

Gaussian processes

- Instead of specifying a prior on Y using a fixed basis of functions and parameters with a Gaussian prior, we can directly specify a prior on Y as a (centered) Gaussian process on the set $\mathcal{X}$.
- A centered Gaussian process on a set $\mathcal{X}$ is entirely determined by its covariance structure, $k(x, x') = \mathbb{E}[G_x G_{x'}]$, which is a positive definite kernel.
- Conversely any pd kernel on $\mathcal{X}$ defines a Gaussian process index by $\mathcal{X}$.
- For any fixed $X_1, \ldots, X_n$ the prior on the values or (Y_i) is Gaussian with covariance matrix K, , $K_{ij} = k(X_i, x_j)$.

- Going back to the regression example, with a GP prior we can directly consider the predictive distribution for a new point (X^*, Y^*): given the design $(X_1, \dots, X_n, X^*)$, the vector $\begin{pmatrix} Y \\ Y^* \end{pmatrix}$ is jointly centered Gaussian with covariance matrix

$$\gamma^2 \begin{pmatrix} K_\sigma & k_* \\ k_*^t & k(X^*, X^*) + \sigma^2 \end{pmatrix},$$

  where $K_\sigma = K + \sigma^2 I_n$, k_* is the column vector with entries $k(X_i, X^*)$;
- hence

$$Y^* | ((X_i, Y_i)_{1 \le i \le n}, X^*) \sim \mathcal{N}(k_*^t K_\sigma^{-1} Y, k(X^*, X^*) + \sigma^2 - k_*^t K_\sigma^{-1} k_*).$$

- (This can be extended to a set of test points)

Going back to the “basis of functions” representation

- From the previous expression we see that we can write the mean of $Y^*|Y$ as a combination $\sum_{1\leq i\leq n}\theta_i k(X_i, X^*)$;
- Thus it seems that we can cast this setting back into the “basis of functions” representation.
- However, there are several caveats in this “equivalence”: first, the prior over the vector Θ is now centered Gaussian with covariance matrix K^{-1}.
- If we want to predict on a new point X^*, we must also include the function $K(X^*, .)$ in the “basis” with a coefficient θ^*. We then get

$$\begin{pmatrix} \Theta \\ \theta^* \end{pmatrix} |((X_i, Y_i)_i, X^*) \sim \mathcal{N}\left(\begin{pmatrix} K_\sigma^{-1}Y \\ 0 \end{pmatrix}, K_*^{-1} - \begin{pmatrix} K_\sigma^{-1} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix}\right)$$

- The point to understand is that with a finite basis of functions of size k, we can only obtain the correct prior marginal on k points. The prior on the coefficients is not “intrinsic”, it depends on the set of points considered. Only the GP is the intrinsic prior representation.

Hyperparameters for GP priors

- Just like before, one can make the GP prior depend on hyperparameters, say α. In this case is can be simply represented as a kernel function k_α depending on parameter.
- These can be once again picked by "evidence maximization" if we can't afford to perform the posterior integration over them (again, generally not).
- In this case the quantity to minimize is

$$G(\alpha) = -\log p(Y|X,\alpha) = \frac{1}{2}\log|K_\alpha| + \frac{1}{2}Y^t K_\alpha^{-1} Y + c\,;$$

- For example, using a simple gradient descent technique, we have

$$\frac{dG}{d\alpha} = \frac{1}{2}\mathrm{tr}\left(K_\alpha^{-1}\frac{dK_\alpha}{d\alpha}\right) - \frac{1}{2}Y^t K_\alpha^{-1}\frac{dK_\alpha}{d\alpha}K_\alpha^{-1}Y\,.$$

- The GP can also optionally include a non-zero mean function that can be similarly estimated.
- The point to remember is that the Bayesian framework offers a methodology for choosing the kernel parameters.

The “Relevance Vector Machine”

- The previous remarks notwithstanding, a common shortcut is to consider a Bayesian model with “data-dependent basis functions” of the form
$$V = \sum_{1 \leq i \leq n} \theta_i k(X_i, X^*)\,;$$
- and consider some fixed explicit prior on the parameters.
- A Gamma prior with different parameter for each θ_i gives rise to a sparse Bayesian model called the “Relevance Vector Machine”.

Gaussian process for classification

- In the case of using a GP prior for classification, we specify a GP prior on the latent variables V.
- The problem is that the posterior on the latent variables is not Gaussian (even with fixed hyperparameters).
- Again, we have to use some approximation. A common approach is to approximate the posterior on V_S (the latent variables corresponding to the observed labels) by a Gaussian $\widetilde{Q}(V_S|X_S, Y_S)$.
- In this case, for the latent variables V^* corresponding to the testing points, the distribution $P(V^*|V_S)$ is given by the Gaussian prior hence the approximate posterior

$$P_{approx.}(V^*|X^*, X_S, Y_S) = \int_{V_S} P(V^*|V_S, X^*)\widetilde{Q}(V_S|X_S, Y_S)dV_S$$

is also Gaussian.

The PAC-Bayes bound

- The PAC-Bayes bound concerns randomized classifiers (using a data-dependent randomization measure Θ_S) and concentrates on the generalization error averaged over the randomization:

$$\mathcal{E}(\Theta) = \mathbb{E}_{f\sim\Theta}\mathbb{E}_{X,Y}\left[\mathbb{1}\{f(X) \neq Y\}\right] ,$$

and the relation to its empirical counterpart $\widehat{\mathcal{E}}(\Theta, S)$.

Theorem

Let Π be a prior distribution on a set of classifiers $\mathcal{F}$. Then with probability $1-\delta$ over the draw of the training set S, for any distribution Θ over $\mathcal{F}$:

$$D_+(\widehat{\mathcal{E}}(\Theta), \mathcal{E}(\Theta)) \leq \frac{1}{m}\left(KL(\Theta, \Pi) + \log(m+1)\delta^{-1}\right) .$$

- NB: compare with Occam's Hammer approach which is not averaged over Θ.

- We can apply the PAC-Bayes bound to the “Gibbs” version of a Bayesian classifier which consists in:
- (1) drawing a latent function $V(.)$ according to (an approximation of) the posterior.
- (2) for a test point X^* , output the sign of $V(X^*)$.
- The Radon-Nikodym derivative of the (approximate) posterior wrt. the prior is given by

$$\frac{\widetilde{Q}(V_S|X_S, Y_S)}{P(V_S|X_S)} .$$

- For a Gaussian prior $P(V_S|X_S) \sim \mathcal{N}(0, K_S)$ and an approximate posterior written as $\widetilde{Q}(V_S|X_S, Y_S) = \mathcal{N}(K_S\widetilde{\alpha}_S, \Sigma_S)$, we have

$$KL(\widetilde{Q}, P) = \frac{1}{2}\left(\log|\Sigma_S^{-1}K_S| + \mathrm{tr}(\Sigma_S^{-1}K_S)^{-1} + \widetilde{\alpha}_S^t K_S \widetilde{\alpha}_S - n\right)$$

Link between Gibbs, Bayesian and voting classifiers

- Consider the majority vote decision by classifiers drawn according to the randomization. Then the error of this voting rule is bounded by twice the averaged (over randomization) error of the Gibbs classifier. (In fact this is true for any fixed training point (X^*, Y^*).
- If the posterior distribution of the latent variable is symmetric around its mean at any test point, then the voting classifier coincides with the Bayesian classifier.