9.520: Class 20

Bayesian Interpretations

Tomaso Poggio and Sayan Mukherjee

Plan

- Bayesian interpretation of Regularization
- Bayesian interpretation of the regularizer
- Bayesian interpretation of quadratic loss
- Bayesian interpretation of SVM loss
- Consistency check of MAP and mean solutions for quadratic loss
- Synthesizing kernels from data: bayesian foundations
- Selection (called "alignment") as a special case of kernel synthesis

Bayesian Interpretation of RN, SVM, and BPD in Regression

Consider

$$\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(\mathbf{x}_i))^2 + \lambda \|f\|_K^2$$

We will show that there is a Bayesian interpretation of RN in which the data term – that is the term with the loss function – is a model of the noise and the stabilizer is a prior on the hypothesis space of functions f.

Definitions

1. $D_\ell = \{(\mathbf{x}_i, y_i)\}$ for $i = 1, \cdots, \ell$ is the set of training examples

2. $\mathcal{P}[f|D_\ell]$ is the conditional probability of the function f given the examples g.

3. $\mathcal{P}[D_\ell|f]$ is the conditional probability of g given f, i.e. a model of the noise.

4. $\mathcal{P}[f]$ is the *a priori* probability of the random field f.

Posterior Probability

The posterior distribution $\mathcal{P}[f|g]$ can be computed by applying Bayes rule:

$$\mathcal{P}[f|D_\ell] = \frac{\mathcal{P}[D_\ell|f]\ \mathcal{P}[f]}{P(D_\ell)}.$$

If the noise is normally distributed with variance σ, then the probability $\mathcal{P}[D_\ell|f]$ is

$$\mathcal{P}[D_\ell|f] = \frac{1}{Z_L} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))^2}$$

where Z_L is a normalization constant.

Posterior Probability

Informally (we will make it precise later), if

$$\mathcal{P}[f] = \frac{1}{Z_r} e^{-\|f\|_K^2}$$

where Z_r is another normalization constant, then

$$\mathcal{P}[f|D_\ell] = \frac{1}{Z_D Z_L Z_r} e^{-\left(\frac{1}{2\sigma^2}\sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))^2 + \|f\|_K^2\right)}$$

MAP Estimate

One of the several possible estimates of f from $\mathcal{P}[f|D_\ell]$ is the so called MAP estimate, that is

$$\max \mathcal{P}[f|D_\ell] = \min \sum_{i=1}^{\ell} (y_i - f(\mathbf{x}_i))^2 + 2\sigma^2 \|f\|_K^2 \ .$$

which is the same as the regularization functional if

$$\lambda = 2\sigma^2/\ell.$$

Bayesian Interpretation of the Data Term (quadratic loss)

As we just showed, the quadratic loss (the standard RN case) corresponds in the Bayesian interpretation to assuming that the data y_i are affected by additive independent Gaussian noise processes, i.e. $y_i = f(x_i) + \epsilon_i$ with $E[\epsilon_j \epsilon_j] = 2\delta_{i,j}$

$$P(\mathbf{y}|f) \propto \exp(-\sum(y_i - f(x_i))^2)$$

Bayesian Interpretation of the Data Term (nonquadratic loss)

To find the Bayesian interpretation of the SVM loss, we now assume a more general form of noise. We assume that the data are affected by additive independent noise sampled form a continuous mixture of Gaussian distributions with variance β and mean μ according to

$$P(\mathbf{y}|f) \propto \exp\left(-\int_0^\infty d\beta \int_{-\infty}^\infty d\mu \sqrt{\beta} e^{-\beta(y-f(x)-\mu)^2} P(\beta,\mu)\right),$$

The previous case of quadratic loss corresponds to

$$P(\beta,\mu) = \delta\left(\beta - \frac{1}{2\sigma^2}\right)\delta(\mu).$$

Bayesian Interpretation of the Data Term (absolute loss)

To find $P(\beta, \mu)$ that yields a given loss function $V(\gamma)$ we have to solve

$$V(\gamma) = -\log \int_0^\infty d\beta \int_{-\infty}^\infty d\mu \sqrt{\beta} e^{-\beta(\gamma-\mu)^2} P(\beta, \mu),$$

where $\gamma = y - f(x)$.

For the absolute loss function $V(\gamma) = |\gamma|$. Then

$$P(\beta, \mu) = \beta^{-2} e^{-\frac{1}{4\beta}} \, \delta(\mu).$$

For unbiased noise distributions the above derivation can be obtained via the inverse Laplace transform.

Bayesian Interpretation of the Data Term (SVM loss)

Consider now the case of the SVM loss function $V_\epsilon(\gamma) = \max\{|\gamma| - \epsilon, 0\}$. To solve for $P_\epsilon(\beta, \mu)$ we assume independence

$$P_\epsilon(\beta, \mu) = P(\beta) P_\epsilon(\mu).$$

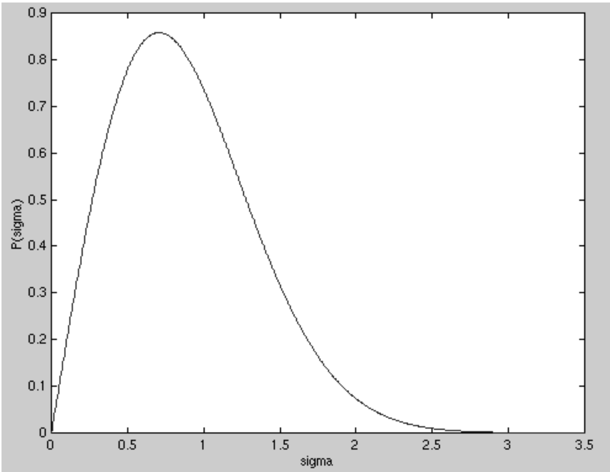
Solving

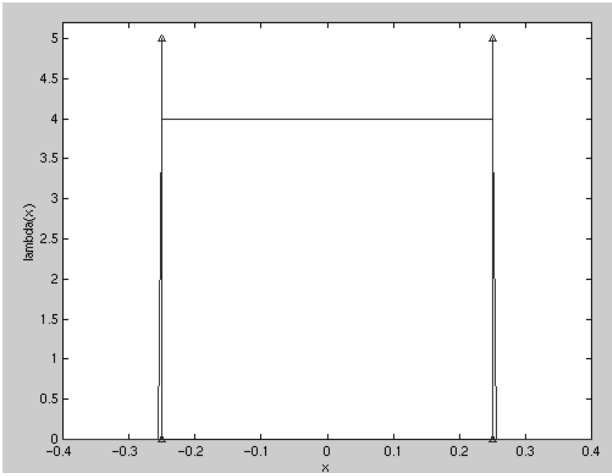
$$V_\epsilon(\gamma) = -\log \int_0^\infty d\beta \int_{-\infty}^\infty d\mu \sqrt{\beta} e^{-\beta(\gamma-\mu)^2} P(\beta) P_\epsilon(\mu)$$

results in

$$\begin{aligned} P(\beta) &= \beta^{-2} e^{-\frac{1}{4\beta}} \,, \\ P_\epsilon(\mu) &= \frac{1}{2(\epsilon+1)} \left(\chi_{[-\epsilon,\epsilon]}(\mu) + \delta(\mu - \epsilon) + \delta(\mu + \epsilon) \right). \end{aligned}$$

Bayesian Interpretation of the Data Term (SVM)





Bayesian Interpretation of the Data Term (SVM loss and absolute loss)

Note $\lim_{\epsilon \to 0} V_\epsilon = |\gamma|$

So

$$P_0(\mu) = \frac{1}{2}\Big(\chi_{[-0,0]}(\mu) + \delta(\mu) + \delta(\mu)\Big) = \delta(\mu)$$

and

$$P(\beta, \mu) = \beta^{-2} e^{-\frac{1}{4\beta}}\, \delta(\mu),$$

as is the case for absolute loss.

Bayesian Interpretation of the Stabilizer

The stabilizer $\|f\|_K^2$ is the same for RN and SVM. Let us consider the corresponding prior in a Bayesian interpretation within the framework of RKHS:

$$P(f) = \frac{1}{Z_r} \exp(-\|f\|_K^2) \propto \exp(-\sum_{n=1}^{\infty} \frac{c_n^2}{\lambda_n}) = \exp(-\mathbf{c}^\top \Lambda^{-1} \mathbf{c}).$$

Thus, the stabilizer can be thought of as measuring a Mahalanobis "norm" with the positive definite matrix Λ playing the role of a (diagonal) covariance matrix. The most likely hypotheses are the ones with small RKHS norm.

Bayesian Interpretation of RN and SVM.

- For SVM the prior is the same Gaussian prior, but the noise model is different and is NOT Gaussian additive as in RN.

- Thus also for SVM (regression) the prior $P(f)$ gives a probability measure to f in terms of the Mahalanobis "norm" or equivalently by the norm in the RKHS defined by R, which is a covariance function (positive definite!)

Why a Bayesian Interpretation can be Misleading

Minimization of functionals such as $H_{RN}(f)$ and $H_{SVM}(f)$ can be interpreted as corresponding to the MAP estimate of the posterior probability of f given the data, for certain models of the noise and for a specific Gaussian prior on the space of functions f.

Notice that a Bayesian interpretation of this type is *inconsistent* with Structural Risk Minimization and more generally with Vapnik's analysis of the learning problem. Let us see why (Vapnik).

Why a Bayesian Interpretation can be Misleading

Consider regularization (including SVM). The Bayesian interpretation with a MAP estimates leads to

$$\min H[f] = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))^2 + \frac{1}{\ell}2\sigma^2\|f\|_K^2 \ .$$

Regularization (in general and as implied by VC theory) corresponds to

$$\min H_{RN}[f] = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))^2 + \lambda\|f\|_K^2 \ .$$

where λ is found by solving the Ivanov problem

$$\min \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - f(\mathbf{x}_i))^2$$

subject to

$$\|f\|_K^2 \leq A$$

Why a Bayesian Interpretation can be Misleading

The parameter λ in regularization and SVM is a function of the data (through the SRM principle) and in particular is $\lambda(\ell)$. In the Bayes interpretation $\tilde{\lambda}$ depends on the data as $\frac{2\sigma^2}{\ell}$: notice that σ has to be part of the prior and therefore has to be independent of the size ℓ of the training data. It seems unlikely that λ could simply depend on $\frac{1}{\ell}$ as the Bayesian interpretation requires for consistency. For instance note that in the statistical interpretation of classical regularization (Ivanov, Tikhonov, Arsenin) the asymptotic dependence of λ on ℓ is different from the one dictated by the Bayesian interpretation. In fact (Vapnik, 1995, 1998)

$$\lim_{\ell \to \infty} \lambda(\ell) = 0$$

$$\lim_{\ell \to \infty} \ell \lambda(\ell) = \infty$$

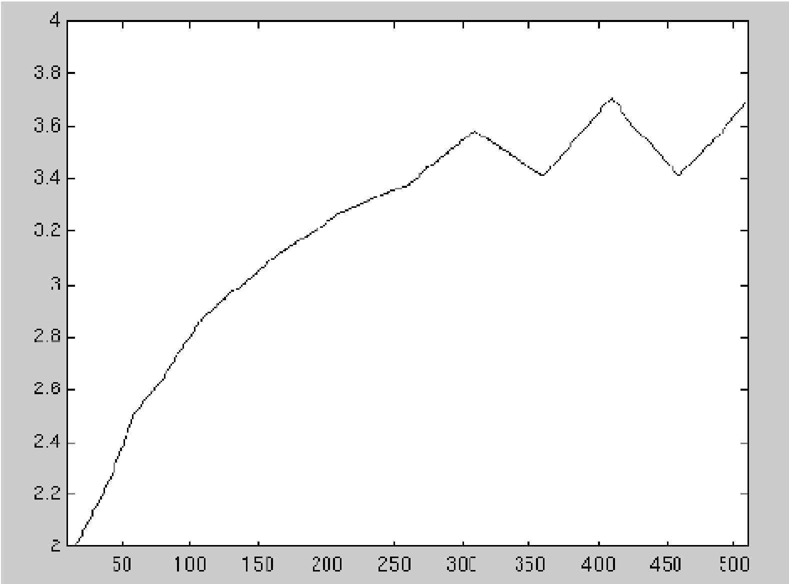
implying a dependence of the type $\lambda(\ell) = O(log\ell/\ell)$. A similar dependence is probably implied by results of Cucker and Smale, 2002. Notice that this is a sufficient and not a necessary condition. Here an interesting question (a project?): which λ dependence does stability imply?

Why a Bayesian Interpretation can be Misleading: another point

The Bayesian interpretation forces one to interpret the loss function in the usual regularization functional (this could be modified but this is another story) as a model of the noise. This seems a somewhat unnatural constraint: one would expect to have a choice of cost independently of the noise type. Conjecture: prove that a probablistic model of the SVMC loss cannot be interpreted in a natural way in terms of a noise model': **project**?

The argument is that $|1 - fy|_+$ cannot be "naturally" interpreted as additive or multiplicative noise. It is a noise that affects real-valued f to give $-1, +1$ with probability that depends on fy. However, we may think of taking $sign(f)$: in this case then the noise flips the true sign with probability ??

From Last Year Class Project...



Consistency check of MAP and mean solutions for quadratic loss (from Pontil-Poggio)

D_ℓ : the set of i.i.d. examples $\{(x_i, y_i) \in X \times Y\}_{i=1}^{\ell}$, etc.

Introduce the new basis functions $\varphi_n = \sqrt{\lambda_n}\phi_n$. A function $f \in \mathcal{H}_K$ has a unique representation, $f = \sum_n b_n \varphi_n$, with $\|f\|_K^2 = \sum_n b_n^2$.

Bayesian Average

$$\bar{f} = \int P(f|D_\ell)d(f) \tag{1}$$

where $P(f|D_\ell) = \frac{P(D_\ell|f)P(f)}{P(D_\ell)}$.

In the φ_n's representation, Eq. 1 can be written as

$$\bar{f} = \mathcal{Z} \int \prod_{n=1}^{\infty} db_n b^T \phi \, \text{exp}\{-H(b)\} \tag{2}$$

with $\mathcal{Z}$ a normalization constant and

$$H(b) = \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i - \sum_{n=1}^{\infty} b_n \varphi_n(x_i)) + \lambda \sum_{n=1}^{\infty} b_n^2.$$

Bayesian Average (cont.)

The integral is not well defined (it's not clear what $\prod_{n=1}^{\infty} dc_n$ means). We define the average function $\bar{f}_N$ and study the limit for N going to infinite afterwards. Thus we define

$$H_N(b) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(y_i - \sum_{n=1}^{N} b_n \varphi_n(x_i) \right)^2 + \lambda \sum_{n=1}^{N} b_n^2$$

$$\bar{f}_N := \mathcal{Z}_N \int \prod_{n=1}^{N} db_n (b^T \varphi) \exp\{-H_N(b)\} \tag{3}$$

Bayesian Average (cont.)

We write:

$$
\begin{aligned}
H_N(b) &= \frac{1}{\ell}\sum_{i=1}^{\ell} y_i^2 - 2\sum_{n=1}^{N} b_n\Big(\frac{1}{\ell}\sum_i \varphi_n(x_i)y_i\Big) + \\
&\quad \sum_{n,m} b_n b_m \frac{1}{\ell}\sum_i \varphi_n(x_i)\varphi_m(x_i) + \lambda\sum_{n=1}^{N} b_n^2 \\
&= \frac{1}{\ell}\sum_i y_i^2 - 2b^T\tilde{y} + b^T(\lambda I + M)b
\end{aligned}
$$

where we have defined $\tilde{y}_n = \frac{1}{\ell}\sum_i \varphi_n(x_i)y_i$ and $M_{nm} = \frac{1}{\ell}\sum_i \varphi_n(x_i)\varphi_m(x_i)$. The integral in Eq. 3 can be rewritten as

$$
\bar{f}_N = \mathcal{Z}_N \exp\{-\frac{1}{\ell}\sum_i y_i^2\}\int \prod_{n=1}^{N} db_n (b'\varphi) \exp\{-b^T(\lambda I + M)b + 2b^T\tilde{y}\} \quad (4)
$$

Bayesian Average (cont.)

Using the appropriate integral in Appendix A we have

$$\bar{f}_N(x) = \sum_{n=1}^{N} \varphi_n(x) \sum_{m=1}^{N} (\lambda I + M)^{-1}_{nm} \tilde{y}_m \tag{5}$$

which is the same at the MAP solution of regularization networks when the kernel function is the truncated series, $K^N(x,t) = \sum_{i=1}^{N} \varphi_n(x)\varphi_n(t)$. We write

$$\bar{f}_N(x) = \sum_{i=1}^{\ell} \alpha_i^N K^N(x_i, x)$$

with

$$\alpha_i^N = \sum_{j=1}^{\ell} (K + \lambda I)^{-1}_{ij} y_j$$

Now study the limit $N \to \infty$. We hope that $\bar{f}_N$ indeed converges to $\bar{f}$ in the RKHS. Then from the property of this space we hope to deduce that the convergence also holds in the norm of $C(X)$. Finishing this proof is a 2003 class project!

Correlation

We compute the variance of the solution:

$$C(x,y) = E\left[\big(f(x) - \bar{f}(x)\big)\big(f(y) - \bar{f}(y)\big)\right] \quad (6)$$

where E denotes the average w.r.t $P(f|D_m)$. Again, we study this quantity as the limit of a well defined one,

$$\begin{aligned} C_N(x,y) &= E\left[\big(f_N(x) - \bar{f}_N(x)\big)\big(f_N(y) - \bar{f}_N(y)\big)\right] \\ &= E\left[f_N(x) f_N(y)\right] + E\left[f_N(x)\right] E\left[f_N(y)\right]. \end{aligned}$$

Using the gaussian integral in Appendix we obtain:

$$C_N(x,y) = \frac{1}{2} \sum_{n,m=1}^{N} \varphi_n(x)(\lambda I + M)^{-1}_{nm} \varphi_m(y)$$

Note that when $\lambda \to \infty$ we get $K_N(x,y)$, so when no data term is present the best guess for the correlation function is just the kernel itself.

A Priori Information and "kernel synthesis"

Consider a special case of the regression-classification problem: in addition to the training data – values of f at locations $\mathbf{x}_i$ – we have information about the hypothesis space that is the class of functions to which f belongs. In particular, we know examples of f in the space and we know or can estimate (in practice often impossible: more later!) the correlation function R. Formally: f belongs to a set of functions f_α with distribution $P(\alpha)$. Then

$$R(\mathbf{x}, \mathbf{y}) = E[(f_\alpha(\mathbf{x}) f_\alpha(\mathbf{y})]$$

where $E[\cdot]$ denotes expectation with respect to $P(\alpha)$. We assume that $E[f_\alpha(\mathbf{x})] = 0$.

Since R is positive definite it induces a RKHS with the λ_n defined by the eigenvalue problem satisfied by R. It follows that we have synthesized a "natural" kernel R – among the many possible – for solving the regression-classification problem from discrete data for f.

Example of R

The $sinc$ function is a translation invariant correlation function associated with the hypothesis space consisting of one-dimensional band-limited functions with a flat Fourier spectrum up to f_c (and zero for higher frequencies). The $sinc$ function is a positive definite reproducing kernel with negative lobes.

Sometime possible Kernel synthesis: regression example

- Assume that the problem is to estimate the image f on a regular grid from sparse data y_i at location $\mathbf{x}_i$; $\mathbf{x} = (x, y)$ on the plane.

- Assume that I have full resolution images of the same type f_α drawn from a probability distribution $P(\alpha)$.

- Remember that in the Bayesian interpretation choosing a kernel K is equivalent to assuming a Gaussian prior on f with covariance equal to K.

- Thus an empirical estimate of the correlation function associated with a function f should be used, *if* it is available, as the kernel. Thus $K(x, y) = E(f_\alpha(x) f_\alpha(y))$.

- The previous assumption is equivalent to assuming that the RKHS is the span of the f_α with the dot product induced by K above.

- Problem, may be a project: Suppose I know that the prior on f is NOT Gaussian. What happens? What can I say?

Usually impossible kernel synthesis: classification

In the classification case, unlike the special regression case described earlier, it is usually *impossible* to obtain an empirical estimate of the correlation function

$$R(\mathbf{x},\mathbf{y}) = E[f_\alpha(\mathbf{x}) f_\alpha(\mathbf{y})]$$

because a) the dimensionality is usually too high and b) R cannot be estimated at "all" x, y (unlike the previous grid case).

Classification: same scenario, another point of view: *RKHS of experts*.

Assume I have a set of examples of functions from the hypothesis space i.e. real-valued classifiers of the same type, say a set of face detection experts or algorithms. Then I consider the RKHS induced by the span of such experts, that is functions $f(\mathbf{x}) = \sum b_\alpha t_\alpha(\mathbf{x})$. The RKHS norm is defined as $|f|_K^2 = \mathbf{b}^T \mathbf{\Sigma}^{-1} \mathbf{b}$, with $\mathbf{\Sigma} = \sum P_\alpha t_\alpha(\mathbf{x}) t_\alpha(\mathbf{y})$ being the correlation function. The $\phi_i(\mathbf{x})$ are linear combinations of the experts t_α; they are orthogonal; they are the solutions of the eigenvalue problem associated with the integral operator induced by Σ that is

$$\int \mathbf{\Sigma}(\mathbf{x}, \mathbf{y}) \phi_i(\mathbf{x}) dx = \lambda_i \phi_i(\mathbf{y}).$$

Classification: same scenario, another point of view

Of course

$$K(\mathbf{x}, \mathbf{y}) = \sum \lambda_j \phi_j(\mathbf{x}) \phi_j(\mathbf{y}) = \Sigma(\mathbf{x}, \mathbf{y})$$

Thus regularization finds in this case the optimal combination of experts with a L^2 stabilizer. There are connections here with Adaboost, but this is another story.

Classification: a different scenario and why alignment may be heretical in the Bayesian church

Assume now that we have a examples of q hypothesis spaces, in the form of a set of experts for each of the q hypothesis spaces. Equivalently we have estimates of the q associated kernels K_m.

What we could do is select the "optimal" kernel K_m by looking at the following score

$$a_m = \frac{(K_m, Y)_F}{||K_m||_F ||Y||_F} = \frac{(K_m, Y)_F}{\ell ||K_m||_F},$$

where the norms and inner products are Frobenious norms ($||X||_F = \sqrt{\sum_{i,j} X_{i,j}^2}$) and the matrix Y has elements $Y_{i,j} = y_i y_j$. So we are selecting a kernel by checking which kernel best "aligns" with the labels.

Classification: a different scenario and why alignment may be heretical in the Bayesian church

From a Bayesian point of view each of the K_m corresponds to a different prior. If we want to do something rather heretical in a strict Bayesian world we could choose the prior that fits our data best. This is exactly what *alignment* does! From a learning theory point of view such an approach may be OK *iff* done in the spirit of SRM – with kernels defining a structure of hypothesis spaces. This would require a change in the alignment process: a new project?

Appendix A: Gaussian Integrals

We state here some basic results (without proofs) on Gaussian integrals. Let $w \in \mathbb{R}^N$, A a $N \times N$ real symmetric matrix which we assume to be strictly positive definite.

$$I(a, A) = \int dw \exp\{-w'Aw + w'a\} = (2\pi)^{\frac{N}{2}} \det(A)^{-\frac{1}{2}} \exp\{\frac{1}{2} a'A^{-1}a\} \quad (7)$$

where the integration is over $\mathbb{R}^N$. Similarly

$$I_u(a, A) = \int dw(w'u) \exp\{-w'Aw + w'a\} = I(a, A)u'A^{-1}u \quad (8)$$

$$\begin{aligned} I_{u,v}(a, A) &= \int dw(w'u)(w'v) \exp\{-w'Aw + w'a\} \\ &= I(a, A) = \left[u'A^{-1}v + (u'A^{-1}a)(v'A^{-1}a)\right] \end{aligned} \quad (9)$$