Statistical Techniques in Robotics (16-831, F14) Lecture#NN (Tuesday November 4)

Kernel methods and Bayesian linear regression

Lecturer: Drew Bagnell *Scribe:Arun Srivatsan* [1]

1 Revisiting Reproducing Hilbert Spaces

Recall form the previous lecture that a function $f \in \mathcal{H}_K$ is a weighted sum of kernels centered at various locations x_i:

$$f(\cdot) = \sum_{i=1}^{N} \alpha_i K(x_i, \cdot),$$

where K must be symmetric: $K(x_i, x_j) = K(x_j, x_i)$. Also kernel K must be positive definite, i.e., if we define $\mathbf{K}_{ij} = K(x_i, x_j)$, then $\mathbf{K}$ must be positive-semidefinite. For two functions $f, g \in \mathcal{H}_K$, we define an inner product over the RKHS $\mathcal{H}_K$ as follows:

$$\langle f, g \rangle = \sum_i \sum_j \alpha_i \beta_j K(x_i, x_j) = \alpha^T \mathbf{K} \beta, \quad \text{where}$$

$$f = \sum_i \alpha_i K(x_i, \cdot)$$

$$g = \sum_j \beta_j K(x_j, \cdot)$$

This now allows us to define a norm (or seminorm) over $\mathcal{H}_K$ as follows:

$$||f||^2 = \langle f, f \rangle$$

$K(\cdot, \cdot)$ is a reproducing kernel of a Hilbert space $\mathcal{H}$ if $\forall f \in \mathcal{H}, f(x) = \langle K(x, \cdot), f(\cdot) \rangle$

2 SVM loss with online Kernel

The loss is given by: $L_t = \max(0, 1 - y_i f(x_i))$. Thus we have:

$$\begin{aligned} \nabla L_t &= 0 \quad \text{if} 1 - y_i f(x_i) < 0 \quad \text{correct by margin} \\ &= -y_i K(x_i, \cdot) \quad \text{else} \quad \text{margin violation} \end{aligned}$$

- Number of kernels within constant factor of total points.
- Does not scale well to very large number of data points
- Kernel methods are good when small data, complicated features
- Linear SVM methods are good when large data and simple features

[1] Some content adapted from previous scribes: Carl Doersch

3 Representer Theorem

Regret $\leq \sqrt{F^2 G^2 T}$, where $F^2 = ||f - f^*||_K$, $G^2 = K(x_i, x_i)$. What K gives the same behavious as linear SVM?
Linear Kernel $K(x,y) = x^T y$. Online learning looks like Bayes rule. Bayes rule as an instance of online learning. Find loss fuction L_t, learning rate α_t such that Gaussiam Weighted Majority gives back Bayes rule. Prior in weighted majority, $w_i = p_i$, where $\sum_i p_i = 1$ and $p_i \geq 0$. $W = \sum_i w_i$ and e^* is some expert and m^* be the number of mistakes that e^* makes and m be the number of mistakes the algorithm makes. Then we have:

$$
\begin{aligned}
&2^{m^*} p^* \leq W \leq \frac{3}{4}^m \\
\Rightarrow &2^{m^*} p^* \leq W \leq \frac{4}{3}^{-} m \\
\Rightarrow &m^* + \log_2 p^* \leq \log_2 W \leq -m \log_2 \frac{4}{3} \\
\Rightarrow &m \leq 2.41 m^* + \log_2 \frac{1}{p^*}
\end{aligned}
$$

4 Bayesian Linear Regression (BLR)

In linear regression, the goal is to predict a continuous outcome variable. In particular, let:

- $\theta =$ parameter vector of the learned model
- $x_t \in R =$ set of features at every timestep, used for prediction
- $y_t \in R =$ true outcome

Then our model is as follows:

$$y_t = \theta x_t + \epsilon_t,$$

where ϵ_t is a noise independent of everythign else. This has the following form $y_T \; N(\theta^T x_i, \sigma^2)$. Thus the likelihood if θ is known is:

$$P(y|x, \theta) = \frac{1}{Z} \exp \frac{\theta x}{2\sigma^2}$$

In BLR, we maintain a distribution over the weight vector θ to represent our beliefs about what θ is likely to be. The math is easiest if we restrict this distribution to be a Gaussian: $\theta \in N(,)$

$$P(\theta) = \frac{1}{Z} \exp \frac{-(\theta - \mu)^T \Sigma^{-1} (\theta - \mu)}{2},$$

where Σ is positive definite. This is called moment parameterization of a Gaussian.

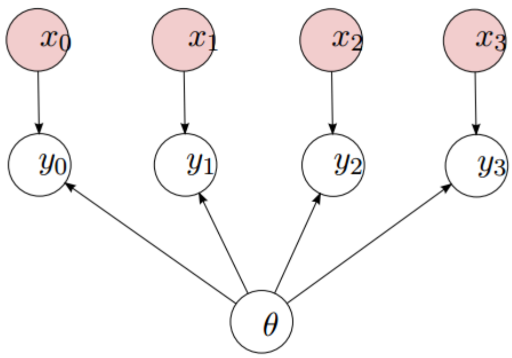

Figure 1: Graphical model of Bayesian Linear Regression

5 Scribed notes

Scribed notes are due a week after the lecture. You may use previous year's notes as a resource, but be thorough and improve upon the existing material. If you adapt a previous scribe's notes, be sure to acknowledge them.

5.1 Instructions

After you have finished scribing your assigned lecture, you should:

- Upload the pdf to the google group
- Send the source documents (.tex and any figures) to the TA
- ...
- Profit

We will assemble all scribed notes to serve as a resource for next year's students.

5.2 Things to change

Before you upload your notes, please remember to change the following:

- Lecture number and your andrewid (file name)
- Lecture number and date (header)
- Lecture topic (header)
- Scribe name (header)
- Any previous scribes (footnote)

Email the TA if you have any questions.