COMP 652 - ECSE 608: Machine Learning - Assignment 2

Posted Wednesday, October 11, 2017
Due Wednesday, November 1 , 2017

---

1. [40 points] **Gaussian Processes**

   For this exercise, you can use either the Python Gaussian Process package from scikit-learn (`sklearn.gaussian_process`) or the GPML toolbox available in Matlab (we encourage you to use the scikit-learn Gaussian Process package). You can also use another programming language but then you need to check whether a GP toolbox is available.

   In order to get a clear idea of what the plots you have to do in this exercise should look like, have a look at the following pages:
   http://scikit-learn.org/stable/auto_examples/gaussian_process/plot_gpr_prior_posterior.html
   http://scikit-learn.org/0.17/auto_examples/gaussian_process/plot_gp_regression.html

   (a) [5 points] The squared exponential covariance function (RBF kernel) is defined by

   $$k(x, x') = v_0 \cdot \exp\left(-\frac{(x - x')^2}{2l^2}\right) \tag{1}$$

   where v_0 is the scale parameter and l is the length parameter. Draw 10 different samples from a GP prior with the RBF kernel with $l = v_0 = 1$ and plot these 10 functions for x ranging from 0 to 5. How does the prior samples change if you use different values for the parameters l and v_0?

   (b) [5 points] Comment on the shape of the functions you plotted previously and try to explain how these properties are related to the definition of the RBF kernel. What do you think will be the effect of the length parameter on the bias-variance trade-off?

   (c) [10 points] We consider the model $y = x \sin x$. Generate a training set of input-output examples (x_i, y_i) of size 50 drawn from this model, where the x_i are drawn uniformly between 0 and 10.

   i. Fit the data using Gaussian Process with a squared exponential covariance function with scale parameters $v_0 = 1.1$, using maximum likelihood estimation of the parameters. Use a starting point for the MLE estimation of the best set of hyper-parameters as 10^{-1}, with a upper bound of 1 and a lower bound of 10^{-3} on the parameters

   For example, this can be done in scikit learn with

   ```
   from sklearn.gaussian_process import GaussianProcessRegressor
   from sklearn.gaussian_process.kernels import RBF

   kernel = 1.1 * RBF(length_scale=0.1, length_scale_bounds=(1e-3, 1.0))
   gp =  GaussianProcessRegressor(kernel=kernel)
   gp.fit(X,y)
   ```

   where the matrix X and vector y are the inputs and outputs from the training set.

   Plot the mean and variance of the GP fit for x ranging from 0 to 15. You should also show the training instances on this plot. What does the variance tell you about the fit of the GP model? If you use a starting point for the length scale l of the RBF kernel to be 10.0, how does the fit of the GP model differ? Comment on the difference.

   ii. Instead of using a RBF kernel, now fit the GP model with a linear kernel (`DotProduct` in scikit learn). Plot the mean and variance of the GP fit. How does the fit of the GP model with this kernel compare to fitting it with an RBF kernel?

iii. Again fit a GP model, this time with a rational quadratic kernel (`RationalQuadratic` in scikit learn) with initial length scale $l = 0.1$. Plot the mean and variance of the GP fit. Comment on the difference using this kernel, compared to using a RBF or linear kernel.

iv. Instead of using a single kernel for the GP model fit, consider using a sum of kernels (in scikit learn, this can simply be done by using the addition operator, e.g. `kernel = DotProduct(...) + RBF(...)`). Fit a GP model with a sum of RBF kernel and a RationalQuadratic kernel. How does the fit of the GP model with a sum of kernels differ, compared to using a single kernel? Plot the mean and variance of the GP model fit. Comment on the difference in your observations.

(d) [10 points] We again consider the model from the preceding question, but this time with noisy observations. Considering the model $y = x \sin x + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0.0.5)$. Generate a training set of input-output examples (x_i, y_i) of size 50 drawn from this model, where the x_i are drawn uniformly between 0 and 10.

i. Fit a GP model for this noisy data using the kernels you considered above in questions $(c.i) - (c.iv)$, i.e. first fit the GP model with a RBF kernel, then with a DotProduct kernel, a RationalQuadratic kernel, and finally with a sum of RBF and RationalQuadratic kernel. How does the fit of the GP model with each of these kernels on a noisy data set differ, compared to what you observed above in (c)? Plot the mean and variance for each of the GP model fit.

ii. A White kernel is often used as part of a sum kernel to account for the noise component of the signal. Tuning the parameter of the White Kernel corresponds to estimating the noise level. For example, in scikit-learn, the White kernel is implemented in `sklearn.gaussian_process.kernels.WhiteKernel`.

For each kernel k you considered above in question $(d.i)$, use the sum of k and a White kernel. For example, for the RBF kernel, now consider a sum of RBF and White kernel. For each of these new kernels, fit a GP model on the data and plot the mean and variance for each of your model fit. What impact do you think the White kernel has on the model fitting? Comment on the difference in your observations.

iii. Report the log marginal likelihood on the training data for each of the kernels in questions (d.i) and (d.ii). Comment on the difference in the log marginal likelihood values. What does the log marginal likelihood value tell you about the fit of the model with these kernels? Explain the differences. Which kernel, or combination of kernels, do you think is good for the GP model fit on this noisy data?

(e) [10 points] Load the Mauna Loa Atmospheric CO2 data. The dataset can be downloaded using `sklearn.datasets.fetch_mldata` and is also available in the `mauna.csv` file. This is a time series of monthly average atmospheric Carbon Dioxide concentrations in parts per million (vol) measured at Mauna Loa in Hawaii. <u>You need to center the data</u> (both input and output). For this question, you will design a kernel that will give a good GP fit for this data.

First plot the data and comment on the trends in this dataset.

Then, you need to find a good kernel or combination of kernels to account for the trends that you observed in the previous plot. You should experiment with the kernels you used in the previous questions (and others if you want) and combinations of these kernels. While you experiment with different kernels, in order to check whether your kernel leads to a good fit of the data, you should

- fit a GP model with the given kernel on the data

- plot the mean and variance of the GP fit for values of x ranging from m to $M + 30$ where m (resp. M) is the minimum (resp maximum) value of the inputs in the training data
- look at the log marginal likelihood on the training data

(no need to include these plots in the report).

Once you have found a good kernel or combinations of kernels, that you think is a good fit to model this CO2 data, report the log marginal likelihood value for your GP model and plot the mean and variance of the model between m and M (defined above). Why do you think the combination of kernels you found is a good kernel for this dataset?

---

2. [10 points] **Graphical Models**

   (a) A doctor gives a patient a (D)rug (drug or no drug) dependent on their (A)ge (old or young) and (G)ender (male or female). Whether or not the patient (R)ecovers (recovers or doesn't recover) depends on all D, A, G. In addition, A is independent of G.

   - Write down and draw the belief network (DAG model) for the above situation.
   - Explain how to compute p(recover | drug)
   - Explain how to compute p(recover | drug, young)

   (b) Consider the factorization of the joint distribution on three variables

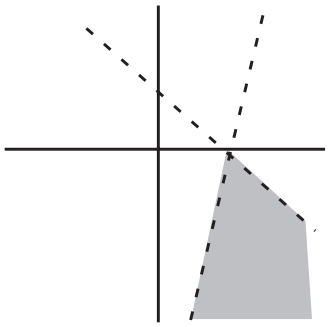
$$p(a, b, c) = p(a|b)p(b|c)p(c) \tag{2}$$

   where all the variables are binary. Draw the graphical model for this distribution. How many parameters are needed to specify this distribution? How many parameters do we need if we make no assumptions on the joint distribution?

---

3. [10 points] **VC dimension**

   Consider the space of points in the plane. Consider the class of hypotheses defined by conjunctions of two linear classifiers (also known as perceptrons), each with two inputs. An example of such a hypothesis is shown in the figure below.



   (a) Show a set of 4 points in the plane that can be shattered by this hypothesis class.

   (b) What is the VC-dimension of this hypothesis class? Explain your answer.

---

4. [20 points] **Markov Random Fields**

   Consider the 2D spin glass model we discussed in class.

   (a) [5 points] Suppose that instead of connecting pixels in a 4-neighborhood, we want to connect them in an 8-neighborhood. Describe what the parameters of the undirected graphical model will be.

(b) [5 points] Suppose that we want to use such a model to capture natural scenes in images. Describe the advantages and disadvantages of this model compared to connecting a pixel only to 4 neighbors.

(c) [10 points] For the 2D Ising model connected as in class, write a Gibbs sampling algorithm, assuming that potentials are represented using linear energy functions and that evidence can be injected along the leftmost edge of the model. Assume the model is an $n \times n$ lattice.

---

5. [15 points] **PCA**

Consider the data set available in the file `hw2pca.txt`; each row represents an instance and the columns represent features. You should split the data into 80% representing the training set and 20% to test the representation. Perform PCA on the data and plot the reconstruction error as a function of the number of dimensions, both on the training set and on the test set, as well as the fraction of the variance accounted for (obtained by looking at the top eigenvalues). Explain what you see and what are the implications for choosing dimensionality of the data.

---

6. [10 points] **PAC learning of concentric circles.**

Let $\mathcal{X} = \mathbb{R}^2$ and consider the set of concepts of the form $c = \{(x, y) : x^2 + y^2 \leq r^2\}$ for some real number r (i.e. the points inside the circle of radius r centered at the origin are labeled $+1$ and the ones outside are labeled -1). Show that this class can be (ε, δ)-PAC-learned from training data of size $m \geq \frac{1}{\varepsilon} \log \frac{1}{\delta}$.