IST 597: Elements of Machine Learning : Midterm (Spring 2021)

Justin Silverman

Instructions:

- Unlike homework assignments, there will be no extensions on this assignment (except in emergency situations).
- Unlike homework assignments, this assignment is to be completed alone, no working with other students.
- You are welcome to use whatever internet / book resources you want here (other than posting these questions on forums like StackExchange and asking someone else to solve the problem for you).
- You have 2 weeks to complete this analysis (this is due to students requesting an extra week), **this is due Thursday Night (March 25th) at midnight**.
- This assignment is worth 20% of your final grade. That equates to roughly 200 homework points.

To assist you, I have added a difficulty indicator to each problem. Here is my translation of those indicators with a *very* approximate time I would expect each problem to take.

* Easy: 1 minute to 1 hour

** Medium: 30 min to 4 hours

*** Difficult: 1 hour to 10 hours

Problem 1 (60 points) ***

In this problem you are going to explore two new areas of machine learning: (1) you are going to learn a bit about time-series analysis, and (2) probabilistic prediction.

In `data_train_midterm_problem1.csv` I have given you historical stock prices for a public company. The variable `stock_price` gives you the price of the stock dating back to the beginning of 2011 and up through the end of the first quarter of 2019. You need to produce a probabilistic forecast of the stock price going up to the end of 2020.

a. Do a brief exploratory data analysis. Notice anything of interest? Keep your responses to less than 1 paragraph and a at most 2-3 figures. One thing you may want to explore is how the data looks after log-transformation. Lots of stock data is more naturally modeled on a log-scale. Of course its your choice, just make sure that in the rest of the problem, if you choose to model on a transformed scale that you give predictions on the original scale of `stock_prices`.

While you may not know time-series analysis, you do know how to use Gaussian Processes regression to learn non-linear functions. What is a time-series but a function over time anyways, right? Lets call the variable `stock_price` or (if you chose to model on log-scale) `log(stock_price)` y_t for notational convenience (e.g., if you are modeling on the log scale $y_t = \log(\text{stock\_price})$ on day t). You are going to be designing a mean function $m(\cdot)$ and kernel function $k(\cdot, \cdot)$ to do this forecasting challenge. But first, you need to think a bit about how to do cross-validation on a time-series.

b. Describe a cross validation approach for this problem. Lets say you want to have 5 folds, for each fold how do you pick which observations are in the training vs. testing set? *Hint: This is less simple than for independent observations, the fact that your measurements are correlated means that it will be non-trivial to design a good cross validation scheme here. Think about what the challenge is:* ***FORECASTING INTO THE FUTURE****.*

c. By choosing $m(\cdot)$ and $K(\cdot, \cdot)$ design and fit the best model you can of the form:

$$y_t \sim N(f(t), \sigma^2)$$
$$f \sim \mathcal{GP}(m(\cdot), K(\cdot))$$

Describe your modeling choices and how you came to them. How did you fit the model? You don't need to report fitted parameter values, just the functional forms of m and K as well as indicating any hyper-parameters that you fit/optimized and how you accomplished that task. (Was there anything you tried that didn't work well?)

So far in this course you have used a number of models to produce point predictions (e.g., forecast the exact label of a test point). Even though, you have been using a number of probabilistic models I have not yet asked you to build probabilistic predictions. These can be incredibly powerful and give you not only a sense of what to expect and how strongly you should expect it. Is our model confident in its predictions or is it unsure? There is no magic bullet here, bad models can be certain that their incorrect predictions are correct. Still, probabilistic predictions and forecasts, when used correctly, can be powerful.

d. Start by visualizing your forecasts. Your output here should be a single plot with multiple things plotted on top of each other (likely will want to make each plot element somewhat transparent so its still readable.) Plot the training data. Now plot the posterior $p(y_*|y)$ evaluated at the training points as well as for the dates specified in `data_test_midterm_problem1.csv`. That

is, predict the stock price at the observed training points and at the unobserved testing points. Its generally hard to plot a distribution that changes over time (no good plotting functions to do this typically), so just plot the posterior mean and the 95% credible interval. The 95% credible interval is the region between the 2.5-th quantile and the 97.5-th quantile of your posterior simulations (e.g., 95% of your posterior samples for y_* should exist within the 95% credible interval). On top of this plot the training data so that you can visualize how the model is fitting the observed data.

e. Now its time to submit your predictions. Rather than just submitting a single prediction (as you have done in prior problems), here you are going to submit 1000 simulated forecasts (*i.e.*, posterior samples). To be clear, I am asking you to submit 1000 samples from the posterior predictive distribution of your model evaluated at the dates specified in `data_test_midterm_problem1.csv` (1000 samples from $p(y_*|y)$). These should be submitted as a comma-separated value text file saved as `[your last name]_forecast_midterm_problem1.csv`. Each row should correspond to a corresponding date in `data_test_midterm_problem1.csv` and each column should be a different sample from the posterior predictive distribution of your model (i.e., the overall file should have `n_test_points` rows and `n_posterior_samples` columns). You will be scored not only on how accurate your posterior mean is to the true stock price, but also whether your probability intervals are calibrated (e.g., even if your mean is wrong, does your 95% credible interval cover the true value tightly and often?).

In the interest of full transparency, I will be using the function `crps_sample` from the R library `scoringRules` using method `edf` to score your forecasts (but don't get bogged down in this detail, I just put it out there just in case you wanted to know).

Problem 2 (40 points) **

You are working for a tech start-up that is trying to develop tools for predicting web traffic for online retailers. Its a high-pressure job, the entire company is based on a completely new data stream that could really revolutionize search optimization and online marketing. If all goes well you and your company can then apply this approach to other web-based marketing problems. Your boss calls and asks you to take a look at the latest data (`data_midterm_problem2.csv`). He has a presentation for potential new clients tomorrow morning and wants you to do a quick analysis on data from 1000 current clients, quantifying how useful your companies data streams (e.g., features; `X1-X13`) are at predicting hourly web-traffic (`y`, number of unique visitors in a 1 hour period).

You ask him to clarify what he means by "useful" and he says: "I just want to know how well these 13 data streams can predict the number of unique visitors. Honestly, I just need something to show these clients. I am sure you have a good idea as to how to go about this. I don't need a visual or anything, I just need a simple statistic that I can report as part of our sales pitch."

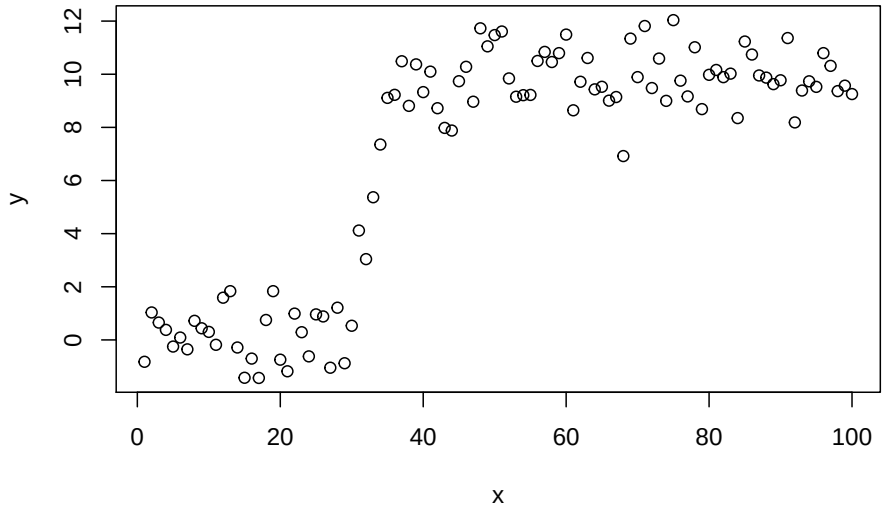
Analyze `data_midterm_problem2.csv`. How do you interpret what your boss is asking for? What do you find and what do you tell him?

Please limit your response to at most about 1 paragraph.

Problem 3 (20 points) *

FYI: There is not a single correct answer to this problem.

Suppose I give you the data in the below figure.



I generated this data using a Gompertz Function with added noise. The Gompertz function is given by:

$$y_i = ae^{-be^{-cx_i}}.$$

a. For N samples $\{(y_1, y_2), \ldots, (y_N, x_N)\}$, write a likelihood model $p(y|x, a, b, c)$ that you expect would provide a reasonable fit for this data. I suggest you plot simulations from your likelihood model to convince yourself whether what you are proposing is/is not a reasonable fit to the data. *(Note, as a training exercise, I am purposefully not giving you the actual data. Learning how to simulate from your model to ascertain whether the model is reflecting what you see visually is a useful skill).*

Now suppose I ask you to fit a Gompertz function to this data, ($y_i = ae^{-be^{-cx_i}}$). The next few questions are designed to make you think carefully about what it means to fit a function to data. I want you to think about what it means to have the "best" estimate for parameters a, b, and c. You are going to create multiple definitions of "best" but don't worry, you don't actually have to solve the problems, just specify them (*i.e.*, write the problem using mathematical notation).

b. Given your likelihood model above, write an optimization problem to estimate a, b, and c using maximum likelihood.
c. Take your likelihood model and extend it to a Bayesian inference problem.
d. Now write this as a penalized likelihood model
e. Now write this as a least squares optimization problem.

Problem 4 (25 points) **

I have hidden an image (made up of 142 two-dimensional data points) in `data_midterm_problem4.csv`. Find the image. Explain what methods you used to find the image. If you don't find the image, describe what you tried.

Problem 5 (20 points) *

Consider the following Ordinary Least Squares estimator for linear regression:

$$\hat{\beta}_{OLS} = \underset{\beta}{\text{argmin}} \sum_i (y_i - \beta^T x_i)^2$$

and the following Bayesian linear regression model (consider σ^2 to be known)

$$y_i \sim N(\beta^T x_i, \sigma^2)$$
$$\beta \sim p(\cdot)$$

for some prior distribution p which can have any arbitrary parameters $(\cdot)$. The *maximum a posteriori* estimate for β given this Bayesian model takes the form:

$$\hat{\beta}_{MAP} = \underset{\beta}{\text{argmax}} \prod_{i=1}^{N} N(y_i|\beta^T x_i, \sigma^2) p(\beta|\cdot).$$

Find a functional form for $p(\beta|\cdot)$ such that $\hat{\beta}_{MAP} = \hat{\beta}_{OLS}$.

Problem 6 (35 points) ***

A friend of yours has designed a small robot to launch a ball from ground-level into the air. He has two dials on the robot, one controls the initial velocity of the ball (really the speed), the other controls the initial angle of the ball. Based on how he wired the robot, he believes that the first knob is related linearly to angle and the second is related linearly to velocity (he is not sure if that information will help you or not). Ultimately, he wants to be able to predict how far the robot will throw the ball as a function of the position of those two knobs, *i.e.*, he wants you to solve a prediction task.

Your friend asks you to build a machine learning model to predict how far the robot will throw the ball (assume all the ground around is flat, e.g., no hills or valleys in the area). He has collected a dataset of 40 observations. For each observation he measures three variables, `y` the range the ball was thrown (in meters), `x1` the value of the first knob `x2` the value of the second knob. He gives you about half of the data in `data_train_midterm_problem6.csv` and asks you to predict the range (`y`) for the other half of the data which he has held out `data_test_midterm_problem6.csv`.

Submit your predictions in a comma-separated values file titled
`[your last name]_predictions_midterm_problem6.csv` (just submit one column that is your predictions for `y`, no column header needed).