CMPS 242: Machine Learning: Fall 2016: HW 2

Due: 25th October 2016

General Instructions (please review carefully):

- The assignment is to be attempted in groups of two. If you choose to not work with a partner, **two** points will be automatically deducted from your score for this homework. It is your responsibility to find a partner.
- Each group needs to submit only one set of solutions. However, each group member should completely understand the group's solutions. It is **strongly** recommended that each student work on all of the problems individually before integrating their solutions into the group consensus. Dividing up the problems so each student does just a subset of them is the wrong approach.
- LaTeX is preferred, but neatly handwritten solutions will also be accepted. All solutions need to be handed over in the class before the beginning of the lectures.
- The names of the group members, and their UCSC ID (@ucsc.edu email address) should prominently be written on the upper left corner of the first page.
- Multiple sheets should be stapled together in the upper **left** corner.
- Solutions to the problems should be clearly labeled with the problem number.
- Although no points are given for neatness, illegible and/or poorly organized solutions can be penalized at the TA discretion.
- Clearly acknowledge sources (web, people, books, etc.), and mention if you discussed the problems with other students or groups. In all cases, the course policy on collaboration applies, and you should refrain from getting direct answers from anybody or any source. If in doubt, please ask the instructors or TAs.

Question 1 (1 + 1 + 0.5 + 0.5 points): Consider the following 1-d dataset with 5 points $X = \{-1, 1, 10, -0.5, 0\}$, on which we are going to perform Gaussian density estimation. For the exercise below, you may use Python for plotting but all the calculations have to be done by hand.

- Compute the Maximum Likelihood Estimate (MLE) of the mean and variance. For the variance, compute both the unbiased and biased versions. Comment on what you observe. In particular, how does the presence of an outlier affect your estimates.
- Assume that you have a $\mathcal{N}(0, 1)$ prior over the mean parameter and set the standard deviation $\sigma^2 = 1$. Compute the posterior distribution of the mean parameter and plot both the prior and the posterior distributions. Comment on what you observe.
- Now suppose we change the prior over the mean parameter to $\mathcal{N}(10, 1)$. Compute the new posterior distribution, plot it, and contrast it with what you observed previously.
- Suppose 2 more data points get added to your dataset:

  $$X = \{-1, 1, 10, -0.5, 0, 2, 0.5\}$$

  Using the same $\mathcal{N}(0, 1)$ prior over the mean parameter, compute and plot the posterior. How does observing new data points affect the posterior?

Question 2 (1 + 0.5 + 2 + 0.5 points): Generate 100 data points as follows: Draw x uniformly at random[1] from $[-100, 100]$. For each x draw t from $\mathcal{N}\{f(x), 1\}$ where $f(x) = 0.1 + 2x + x^2 + 3x^3$. In order to fit this curve, we will make use of the following probabilistic model:

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right)$$

where $y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$.

- Perform MLE estimation of $\mathbf{w}$ and β. You may use the `optimize` module from `scipy` for this task. Comment on how well $\mathbf{w}$ and β match the true parameters used to generate the data. How do the estimates change when you use 1000 or 10,000 data points for your estimates?
- Now suppose that you use $y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$ and repeat the above task. Comment on what changed.
- Refer to the slides from the class, where we added a prior over $\mathbf{w}$ in order to derive Bayesian linear regression. Assume that we set the hyperparameter $\alpha = 1$ and plot the Bayesian estimate of the curve and the uncertainty around the estimate. How well does it match the observed data. How does the estimate change when you use 1000 or 10,000 data points?

[1] Do not forget to seed your random number generators!

- Perform Bayesian linear regression but now set $\alpha = 100$. How does this change your estimates? Again, what happens when you use 1000 or 10,000 data points?

Question 3 (1 + 1 + 0.5 + 0.5 points): Let us assume the following generative model for the data:

$$p(\mathbf{x}|\mathcal{C}_k) = \prod_{i=1}^{D} \mathcal{N}\left(x_i|\mu_{ki}, \sigma_k^2\right)$$

In other words, each coordinate is independently drawn from a 1-d Gaussian distribution, where the variance of the coordinates for a given class k are fixed (σ_k^2), and known.

- Derive the MLE estimate for μ_{ki}.
- Assume that k (the number of classes) is equal to 2, and derive the decision rule of the Naive Bayes classifier.
- Suppose the misclassification costs are not equal. In particular, if the cost of misclassifying class 1 is 10 times more than the cost of misclassifying class 2, how does the decision rule change?
- How does the decision rule change when $k = 3$?

Learning Outcomes After this homework you should

- understand the subtle differences between MLE, MAP, and Bayesian estimation procedures
- become comfortable working with the Gaussian distribution both in the univariate and multivariate forms
- design and develop naive Bayes, a classifier which is used in classifying spam and ham emails
- design a develop logistic regression, a classifier which is used in a variety of applications in the industry

Changelog

- 11 October 2016: First version created.