CSE517A – Homework 2

M. Neumann

14 Feb 2025

- Please keep your written answers brief and to the point. Incorrect or rambling statements can hurt your score on a question.
- If your hand writing is not readable, we **cannot give you credit**. We recommend you type your solutions in LaTeX and compile a .pdf for each answer. **Start every problem on a new page!**
- This will be due **14 Feb 2025** at **11:59pm** with an automatic 3-day extension (cf. course syllabus for more information).
- You may work in groups of at most 2 students.
- Submission instructions:
  - Start every problem on a **new page**.
  - Submissions will be exclusively accepted via **Gradescope**. Find instructions on how to get your Gradescope account and submit your work on the course webpage.

Note, that if you use *any* **resources** outside the course materials to derive (part of) your solution, you will need to cite the source in your homework submission. This also holds for **online sources**. If you collaborate with anyone other than your partner, it is your responsibility to indicate this in your submission. Course materials are the lecture notes, course books, and resources linked therein or on Canvas, plus course materials of any prerequisite course officially listed as such in the course listing.

Citing the source(s) does **not** legitimate the *copying* of existing solutions to any given problem, neither does it legitimate that another person (that is not you or your partner) directly solves any (part of the) problems for you. Please, refer to the **course syllabus** for more details.

Problem 1 (*20 points*) Maximum Likelihood Estimation

For each of the following distributions assume we have observed n draws $x_1, \ldots, x_n$, with $x_i \in \mathbb{R}$. State the log-likelihood function and estimate the parameters for each distribution using maximum likelihood estimation (MLE). You will find that many of these answers are highly intuitive.

(a) (*6 pts*) Estimate μ and σ^2 for the Gaussian distribution with PDF

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

(b) (*7 pts*) Estimate λ for the exponential distribution with probability density function (PDF)

$$f(x; \lambda) = \lambda e^{-\lambda x}, \text{ where } x \geq 0.$$

(c) (*7 pts*) Estimate p for the geometric distribution with probability mass function (PMF)

$$f(x; p) = p(1-p)^x, \text{ where } x \geq 0.$$

Problem 2 (*35 points*) Logistic Regression revisited

(a) (*10 pts*) Recap the logistic regression (LR) model:

$$p(y \mid \mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp\left(-y\mathbf{w}^\top \mathbf{x}\right)} \tag{1}$$

Another way of looking at the LR model is to view it as a Bernoulli distribution:

$$p(y \mid \mathbf{x}, \mathbf{w}) = \text{Ber}(y \mid \text{sigm}(\mathbf{w}^\top \mathbf{x})) \tag{2}$$

where $\text{sigm}(a) = \frac{1}{1+e^{-a}}$ is the *sigmoid* or *logistic* function. Show that the Eq. (1) and Eq. (2) are equivalent representations for the case of binary classification.

(b) (*10 pts*) In order to reduce overfitting we can compute an estimate for LR through the Maximum-A-Posteriori (MAP) estimate:

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\max p(\mathbf{w} \mid X, \mathbf{y})$$

Derive the *negative log posterior* (*nlp*) using Eq. (1) and the following assumption on the prior distribution:

$$\mathbf{w} \sim N(\mathbf{0}, \sigma^2 I)$$
$$p(\mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\mathbf{w}^\top \mathbf{w}}{2\sigma^2}}$$

(c) (*5 pts*) Take a close look at this *negative log posterior*. It should look very familiar. Describe the equivalent SRM formulation (i.e., name the loss function and regularizer) and state the

exact relationship between λ and σ.

(d) (*10 pts*) One benefit of LR is that it is easy to interpret. This can be seen by looking at the log odds:

$$\log \frac{p(y = 1 \mid \mathbf{x}, \mathbf{w})}{p(y = -1 \mid \mathbf{x}, \mathbf{w})}$$

Show that

$$\log \frac{p(y = 1 \mid \mathbf{x}, \mathbf{w})}{p(y = -1 \mid \mathbf{x}, \mathbf{w})} = \mathbf{w}^\top \mathbf{x}$$

Then, give a verbose description of the effect that a change of a feature value has on the LR prediction.

Problem 3 (*25 points*) Bernoulli Naïve Bayes

In this problem, we explore why the naïve assumption is necessary.

After learning about using Bayes rule to make predictions (without the naïve assumption), you want to apply the method to a complex problem. Keep in mind that you want to use the rule:

$$p(Y \mid X) = \frac{p(X \mid Y) \cdot p(Y)}{p(X)}$$

and you want to estimate the parameters of $p(X \mid Y)$ and $p(Y)$. However, before applying the method to your problem, you want to apply to a toy problem first.

(a) (*5 pts*) In the toy problem, $X = [X_1, X_2]$ (so $d = 2$), where X_α is binary. Y is also binary. You want to estimate $p(X \mid Y)$ without the Naïve Bayes assumption, that is you **cannot** write

$$p(X \mid Y) = \prod_{\alpha=1}^{d} p(X_\alpha = x_\alpha \mid Y = y),$$

instead, you must estimate

$$\Pr(X \mid Y) = p(X_1 = x_1, \ldots, X_d = x_d \mid Y = y)$$

for all combinations of the values $x_1, \ldots, x_d$, and y. How many parameters do you have to estimate of for your toy problem?
(Here parameters refers to the estimate of $p(X \mid Y) = p(X_1 = x_1, \ldots, X_d = x_d \mid Y = y)$, and $p(Y = y)$ for some combination of x_α's and y. In our case where $d = 2$, examples of such parameters are $p(X_1 = 1, X_2 = 0 \mid Y = +1)$ and $p(Y = -1)$.)

(b) (*5 pts*) After running the decision rule on your toy problem, you decide to apply it to the actual problem. However, in your problem, $d = 100$. How many parameters do you have to estimate now?

(c) (*5 pts*) When is it necessary for us to make the naïve assumption? Explain by showing how the assumption will affect one of the answers from above.

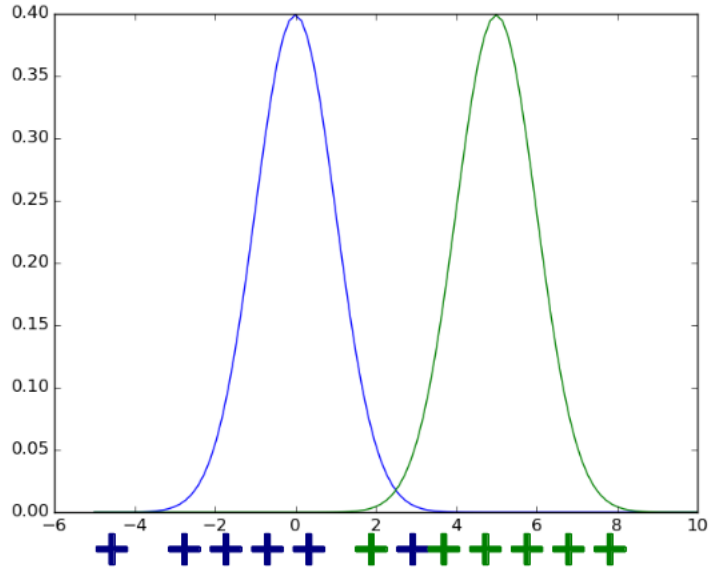
(d) (*10 pts*) Write the Bernoulli naïve Bayes classifier's decision rule as a linear rule. Assuming that we have a binary classification problem and parameter estimates for all the appropriate probabilities and none of them are zero, show that

$$\begin{aligned} h(\mathbf{x}) &= \arg\max_c \; p(y = c) \prod_{\alpha=1}^{d} p(x_\alpha \mid y = c) \\ &= \text{sign}(\mathbf{w}_{\text{BER}}^\top \mathbf{x} + b_{\text{BER}}) \end{aligned}$$

Explicitly provide $\mathbf{w}_{\text{BER}}$ and b_{BER} in terms of the parameter estimates for $p(y = c)$ ($\hat{\pi}_{+1}$ and $\hat{\pi}_{-1}$) and for $p(x_\alpha \mid y = c)$ ($\hat{\theta}_{\alpha,+1}$ and $\hat{\theta}_{\alpha,-1}$).

Problem 4 (*20 points*) Gaussian Naïve Bayes

Previous problems considered only those cases where X consists of discrete values. Now, we will also consider the case where X can take continuous values. We will stick with a binary classification problem $y \in \{-1, +1\}$. Now, we need a different formulation for $p(X_\alpha \mid Y)$. One way is to assume that, for each discrete y, each X_α comes from a Gaussian.



For example, consider the simplified case above, where X_α takes a continuous value (a value along the x-axis) and Y can be either *blue* or *green*. As shown above, for each discrete value of Y (*blue* or *green*), X_α is a random variable from a Gaussian specific to X_α (not some other X_β) and the value of Y. As a result, we see two Gaussians, each generating *blue* data points or *green* data points.

(a) (*10 pts*) State the parameter estimates of the Gaussian naïve Bayes classifier for $p(Y)$ and $p(X_\alpha \mid Y)$. What is the total number of parameters for the Gaussian naïve Bayes classifier?

(b) (*10 pts*) Assuming that each σ only depends on the feature and not the label c (so $\sigma_{\alpha,c} = \sigma_\alpha$), show that Gaussian naïve Bayes is a linear classifier.

In other words, provide the weights w_0 and $\mathbf{w}$ in terms of the parameter estimates $\hat{\pi}_c$, $\hat{\mu}_{\alpha,c}$, and $\hat{\sigma}_\alpha$ by working out the following equation:

$$p(Y = +1 \mid X = \mathbf{x}) = \frac{p(X = \mathbf{x} \mid Y = +1) \cdot \Pr(Y = +1)}{p(X = \mathbf{x})} = \frac{1}{1 + \exp[-(w_0 + \sum_{\alpha=1}^{d} w_\alpha [\mathbf{x}]_\alpha)]}.$$

Note: The result you got is precisely the formulation for logistic regression. However, Gaussian naïve Bayes and logistic regression are estimating different parameters and hence, are different learning algorithms. They only output the same model asymptotically (under special conditions, cf. our assumptions on σ above.).