

Problem 1 (multiclass logistic regression)

Part 1:

- a. Write out the log likelihood L of data $(x_1, y_1), \dots, (x_n, y_n)$ using an i.i.d. assumption.

First, we observe that while $p(y|x, w_1, \dots, w_n)$ is a product of the likelihood of each w , the indicator effectively cancels out every $w_i, i \neq c$ as $z^{\mathbb{1}(y=c)} = z^0 = 1$, where c is the .

With that in mind, we can cast the joint probability of the data as:

$$L = \prod_{i=1}^n \left(\frac{e^{x_i^T w_i}}{\sum_{j=1}^k e^{x_i^T w_j}} \right)^{\mathbb{1}(y=i)}$$

Taking the natural log of this, we have:

$$\ln L = \sum_{i=1}^n \mathbb{1}(y=i) \ln \left(\frac{e^{x_i^T w_i}}{\sum_{j=1}^k e^{x_i^T w_j}} \right)$$

$$\ln L = \sum_{i=1}^n \mathbb{1}(y=i) (x_i^T w_i - \ln \sum_{j=1}^k e^{x_i^T w_j})$$

- b. Calculate $\nabla_{w_i} L$ and $\nabla_{w_i}^2 L$.

Beginning with the log likelihood:

$$\ln L = \sum_{i=1}^n \mathbb{1}(y=i) (x_i^T w_i - \ln \sum_{j=1}^k e^{x_i^T w_j})$$

$$\nabla_{w_i} \ln L = \sum_{i=1}^n \mathbb{1}(y=i) \left(x_i - \frac{1}{\sum_{j=1}^k e^{x_i^T w_j}} (e^{x_i^T w_i}) x_i \right)$$

$$\nabla_{w_i} \ln L = \sum_{i=1}^n \mathbb{1}(y=i) (x_i) \left(1 - \frac{e^{x_i^T w_i}}{\sum_{j=1}^k e^{x_i^T w_j}} \right)$$

ALT:

$$\nabla_{w_i} \ln L = \sum_{i=1}^n (x_i) \left(1 - \frac{e^{x_i^T w_i}}{\sum_{j=1}^k e^{x_i^T w_j}} \right)$$

The numerator $e^{x_i^T w_i}$ makes sense because we take the gradient with respect to w_i . Therefore, the terms of the summation are 0 for every $w_j, w_j \neq w_i$. To interpret the gradient, we see how the greater the probability of x_i , the smaller $\nabla_{w_i} \ln L$, or the less of an impact a change in w_i will have on the likelihood.

Problem 2 (Gaussian kernels)

- a. Use Bayes rule to derive the posterior distribution of λ and identify the name of this distribution.

By Bayes rule, we know that:

$$P(\lambda|x) = \frac{P(x|\lambda)P(\lambda)}{P(x)}$$

Given that we are interested in λ , we can discard the denominator (which ultimately does not involve λ) and rewrite this equation as:

$$P(\lambda|x) \propto P(x|\lambda)P(\lambda)$$

We can then expand the right side of the equation into the joint distribution of $P(x, \lambda)$:

$$\begin{aligned} P(\lambda|x) &\propto \left(\prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right) \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1-1} e^{-\beta_1 \lambda} \\ &\propto \left(\frac{\lambda^{\sum x_i}}{\prod x_i!} e^{-n\lambda} \right) \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1-1} e^{-\beta_1 \lambda} \\ &\propto \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1) \prod x_i!} \lambda^{(\sum x_i + \alpha_1) - 1} e^{-\lambda(n + \beta_1)} \end{aligned}$$

From this we have enough to learn the parameters of the posterior Gamma distribution: $\alpha_2 = \sum x_i + \alpha_1$ and $\beta_2 = n + \beta_1$. This result is still not the complete distribution, however: the fraction is not in the correct form for a Gamma, nor have we addressed the denominator which we dropped earlier. We know, however, that the final distribution must integrate to 1, which allows us to derive the correct values for the other terms (although we will not do so here).

- b. What is the mean and variance of λ under this posterior? Discuss how this relates to your solution to Part 2 of Problem 1.

The posterior λ is $\text{Gamma}(\sum x_i + \alpha_1, n + \beta_1)$

As such, the mean of $\lambda = \frac{\alpha_2}{\beta_2} = \frac{\sum x_i + \alpha_1}{n + \beta_1}$.

The variance is $\frac{\alpha_2}{\beta_2^2} = \frac{\sum x_i + \alpha_1}{(n + \beta_1)^2}$.

In 1.2, we saw that: $\frac{\sum x_i}{n} = \hat{\lambda}_{ML}$.

There is a clear parallel to the Bayesian posterior mean: $\frac{\sum x_i + \alpha_1}{n + \beta_1} = E(\lambda)$, with the difference being the influence of the prior parameters α_1 and β_1 . This relationship represents the Bayesian insight by which our posterior distribution is determined by an interaction between our prior beliefs (α_1, β_1) and our data $(\sum x_i, n)$.

K	Acc
1	.948
2	.93
3	.918
4	.902
5	.894

Table 1: Prediction accuracy for KNN

Problem 3 (Classification)

Part 1

Part 2

- a. In a table, print the mean and standard deviation of the RMSE as a function of p . Using these numbers argue for which value of p is the best.

As this table shows, setting $p = 3$ gives both the lowest RMSE and the lowest variance for any value of p .

p	Mean	Std
1	3.44030468834	0.672995051589
2	2.75919186824	0.623061482354
3	2.64085582004	0.602524538829
4	2.73039848277	0.697498334382

Table 2: RMSE by p

- b. For each value of p , collect $y_{test} - y_{pred}$ for each test example. (Observe that this number can be negative, and there are 20 x 1000 in total.) Plot a histogram of these errors for each p .
- c. For each p , use maximum likelihood to fit a univariate Gaussian to the 20,000 errors from Part 2(b). Describe how you calculated the maximum likelihood values for the mean and variance (this is a univariate case of what we did in class, so no need to re-derive it). What is the log likelihood of these empirical errors using the maximum likelihood values for the mean and variance? Show this as a function of p and discuss how this agrees/disagrees with your conclusion in Part 2(a). What assumptions are best satisfied by the optimal value of p using this approach?

To fit the Gaussian, we calculated both $\hat{\mu}_{ML}$ and $\hat{\sigma}_{ML}^2$ using the standard maximum-likelihood estimators for these parameters: $\frac{1}{n} \sum_{i=1}^n x_i$ and $\frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu}_{ML})^2$.

Looking at this table, $p = 3$ again appears to create the strongest model. The variance (although not the mean) of the error is lowest with this model. Additionally, the log likelihood of this error is the greatest, although by a small margin, compared to all other models. This supports our conclusions in 2(a), where we identified $p = 3$ as being the strongest model.

p	Mean	Var	Log Likelihood
1	0.0163194355852	3.50549766331	-53464.9206058
2	-0.0285981551706	2.82852246649	-49173.3602374
3	0.0409854631276	2.70840898169	-48305.4980812
4	0.0593496182852	2.81743994052	-49094.8436665

Table 3: Log Likelihood of error by p