5. Temperature

Extension Note: Project 2 due date has been extended by 2 days to **July 18 23:59UTC** (Note the UTC time zone).

We will now explore the effects of the temperature parameter in our algorithm.

You will be working in the files `part1/main.py` and `part1/softmax.py` in this problem

Effects of Adjusting Temperature

0.0/1.0 point (graded)

Explain how the temperature parameter affects the probability of a sample $x^{(i)}$ being assigned a label that has a large θ. What about a small θ?

- [x] Larger temperature leads to less variance
- [ ] Smaller temperature leads to less variance ✔
- [x] Smaller temperature makes the distribution more uniform

✖

和正态分布的有点像？

Solution:

Smaller temperature parameter means that there is less variance in our distribution, and larger temperature, more variance. In other words smaller temperature parameter favors larger thetas, and larger temperature parameter makes the distribution more uniform.

Submit You have used 3 of 3 attempts

ℹ Answers are displayed within the problem

Reporting Error Rates

2.0/2.0 points (graded)

Set the temperature parameter to be 0.5, 1, and 2; re-run `run_softmax_on_MNIST` for each one of these (add your code to the specified part in **main.py**).

$\text{Error}|_{T=0.5}$ = 0.084 ✔ **Answer:** 0.084

$\text{Error}|_{T=1}$ = 0.1005 ✔ **Answer:** 0.1005

$\text{Error}|_{T=2}$ = 0.1261 ✔ **Answer:** 0.1261

Submit You have used 1 of 20 attempts

ℹ Answers are displayed within the problem

Return the `temp_parameter` to be 1 before moving on to the next section.

Discussion

Hide Discussion

Topic: Unit 2 Nonlinear Classification, Linear regression, Collaborative Filtering (2 weeks):Project 2: Digit recognition (Part 1) / 5. Temperature

Add a Post

All Posts

My interpretation temperature parameter in question 1

question posted 2 days ago by **khanhedx**

Effect of τ on "variance"

With the formula for predicted probabilities for each label $j \in \{0, \dots, 9\}$ (given a data point x):

$$h(x) = \frac{1}{\sum_{j=0}^{k-1} e^{\theta_j \cdot x/\tau}} \begin{bmatrix} e^{\theta_0 \cdot x/\tau} \\ e^{\theta_1 \cdot x/\tau} \\ \vdots \\ e^{\theta_{k-1} x/\tau} \end{bmatrix}$$

we can see that $e^{\theta_j \cdot x/\tau} = \left(e^{\theta_j \cdot x}\right)^{\frac{1}{\tau}}$. Therefore, as τ gets larger, each $\left(e^{\theta_j \cdot x}\right)^{\frac{1}{\tau}}$ is further squeezed towards zero.

However, one effect of large τ is that the labels with large θ's will be squeezed more than the labels with small theta. For example, given a small θ_1 such that $e^{\theta_1 \cdot x} = 2$, and a large θ_2 such that $e^{\theta_2 \cdot x} = 1000$.

- **For a small τ ($\tau = 3$):** the smaller θ_1 is squeezed by a factor of $\frac{2}{2^{1/3}} = 2^{2/3} = 1.6$ times, while the larger θ_2 is squeezed by a factor of $\frac{10}{10^{1/3}} = 10^{2/3} = 4.6$ times.

- **For a large τ ($\tau = 9$):** the smaller θ_1 is squeezed by a factor of $\frac{2}{2^{1/9}} = 2^{8/9} = 1.85$ times, while the larger θ_2 is squeezed by a factor of $\frac{10}{10^{1/9}} = 10^{8/9} = 7.74$ times.

From the small example above, it can be seen that a large τ will squeeze the $e^{\theta_j \cdot x/\tau}$ terms very hard toward zero, and this squeezing is more pronounced for larger θ. This will in effect cause these terms to not only be small, but also very close to one another in value. As a result, the predicted probabilities will be very similar across the different labels.

An extreme value of large τ is $\tau = \infty$, in which case all $e^{\theta_j \cdot x/\tau}$ terms become $e^0 = 1$, and the predicted probabilities for each label become the same: $\frac{1}{\sum 1} = \frac{1}{10} = 0.1$. Therefore, a large τ will in effect "spread out" the predicted probabilities to be similar across labels, and not concentrated on a few labels with high predicted probabilities, and the rest with low predicted probabilities.

The use of the word "variance"

The above observation could be rephrased as "Larger temperature leads to larger variance" (due to the "spreading out" of predicted probabilities), hence the reverse statement for the answer of the first question: "Smaller temperature leads to smaller variance".

However, I'd argue that the use of the word *variance* is not correct, since the predicted probabilities for each label are separate (Bernoulli) random variables, and not a single random variable with some inherent variance. For one, their generative processes are clearly different, as the predicted probabilities for each label are calculated using different θ_j's, whose distributions are likely to be entirely different.

Another explanation is that there is no such thing as variance of predicted probabilities across labels, as the labels themselves are not real numbers (you can think of the labels being "A" to "J" instead of "0" to "9"). Therefore, there can only be variance for *each* label's predicted probability, but not variance of predicted probabilities *across* labels.

As mentioned earlier, as $\tau = \infty$, the predicted probabilities will become the same across labels (0.1 each). However, this is not a uniform distribution (whose support is the real line), but rather a multinomial distribution with the same probabilities across different events. Interestingly, in that case, the variance for *each* predicted probability will become zero as we are guaranteed that each probability is 0.1, which is the opposite conclusion of the provided answer.

Therefore, I would suggest removing the word "variance" from the question, and replacing it with something like "Larger temperature causes the predicted probabilities to be similar", which has the benefit of being more accurate but also less confusing than "variance" (as multiple discussion threads on this term have shown).

Effect of τ on θ

From the gradient descent formula to update θ:

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta) = \theta + \alpha \frac{1}{\tau n} \sum_{i=1}^{n} \left[x^{(i)} \left([[y^{(i)} == m]] - p(y^{(i)} = m | x^{(i)}, \theta) \right) \right] - \alpha \lambda \theta$$

we can see that smaller τ will cause the first term (the positive "update" term) of the gradient descent formula to be larger compared to the second term (the negative "regularization" term that reduces θ to $\vec{0}$). In other words, a smaller τ will lead to a θ with a larger norm after training.

From my understanding, this effect of τ on θ is unrelated to the effect of τ on "variance" mentioned earlier. Therefore, I would suggest splitting these two effects up into 2 separate questions to avoid confusion.

Of course, please let me know if I have misunderstood anything. Thank you.

This post is visible to everyone.

Add a Response

1 response

khanhedx
2 days ago

One question I still have for the effect on "variance" is that this effect seems to only hold for \(\tau\geq1\). This is because for a small \(\tau\) that is smaller than 1 (say \(\tau=0.02\)), we have \((e^{\theta _ j \cdot x })^\frac{1}{\tau}=(e^{\theta _ j \cdot x })^\frac{1}{0.02}=(e^{\theta _ j \cdot x })^{50}\).

Therefore, it seems that in this case, large \(\theta\)'s (hence large \(e^{\theta _ j \cdot x }\)) are magnified even further by the power of \(50\). This leads to larger separations between predicted probabilities, and thus a larger "variance", which is the opposite of the provided answer.

Please let me know if my understanding is correct or not. Thank you.

Add a comment

Showing all responses

Add a response:

Preview

Submit