Cross-Entropy Loss

Natalie Parde
UIC CS 421

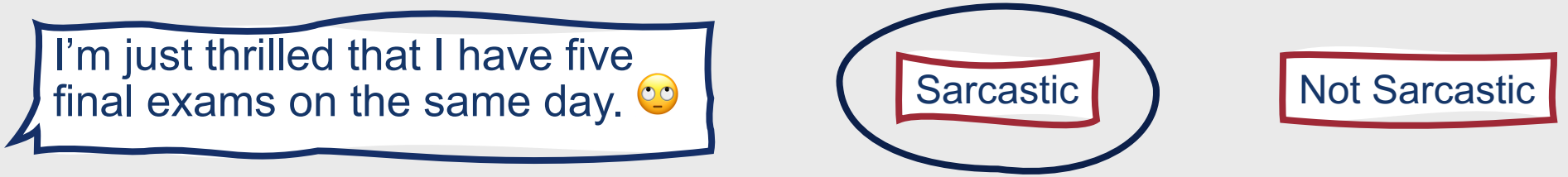
Loss Function

- We need to determine the distance between the predicted and true output value
  - How much does $\hat{y}$ differ from y?
- We do this using a **conditional maximum likelihood estimation**
  - Select *w* and *b* such that they maximize the log probability of the true *y* values in the training data, given their observations *x*
- This results in a **negative log likelihood loss**
  - More commonly referred to as **cross-entropy loss**

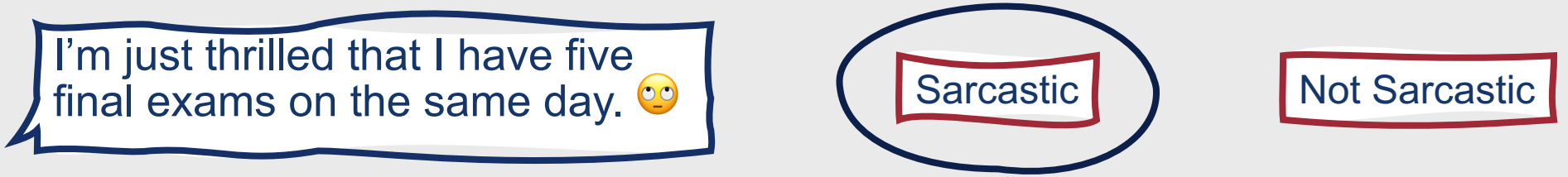
Cross-Entropy Loss

- Most common loss function for many classification tasks
- Measures the distance between the probability distributions of predicted and actual values
  - $loss(y_i, \widehat{y_i}) = -\sum_{c=1}^{|C|} p_{i,c} \log \widehat{p_{i,c}}$
    - *C* is the set of all possible classes
    - $p_{i,c}$ is the actual probability that instance *i* should be labeled with class *c*
    - $\widehat{p_{i,c}}$ is the predicted probability that instance *i* should be labeled with class *c*
- Observations with a big distance between the predicted and actual values have much higher cross-entropy loss than observations with only a small distance between the two values

Example: Cross-Entropy Loss

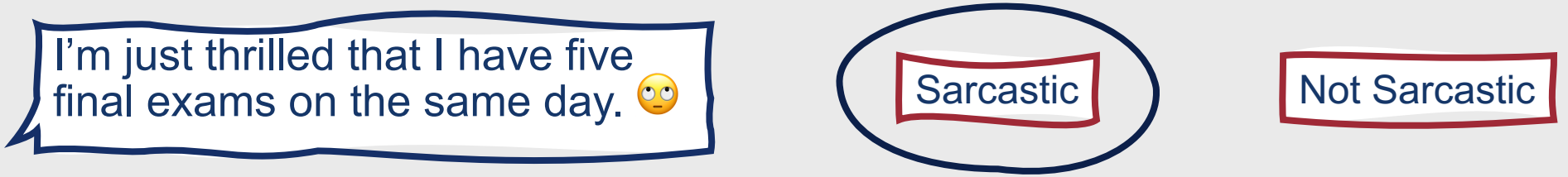


Example: Cross-Entropy Loss



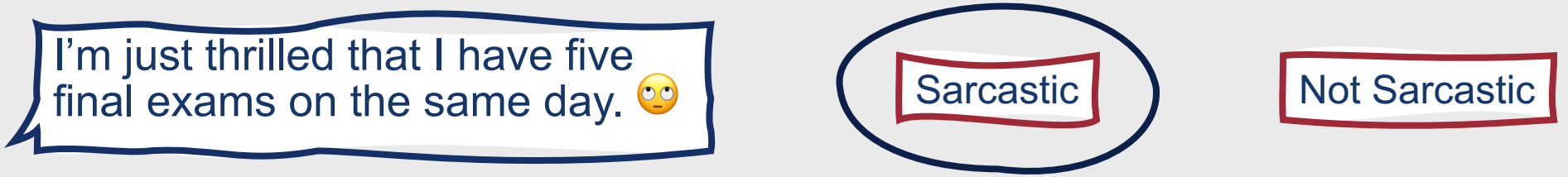
Instance	Predicted Probability: Sarcastic	Predicted Probability: Not Sarcastic	Actual Probability: Sarcastic	Actual Probability: Not Sarcastic
I'm just thrilled that I have five final exams on the same day. 🙄			1	0

Example: Cross-Entropy Loss



Instance	Predicted Probability: Sarcastic	Predicted Probability: Not Sarcastic	Actual Probability: Sarcastic	Actual Probability: Not Sarcastic
I'm just thrilled that I have five final exams on the same day. 🙄	0.96	0.04	1	0

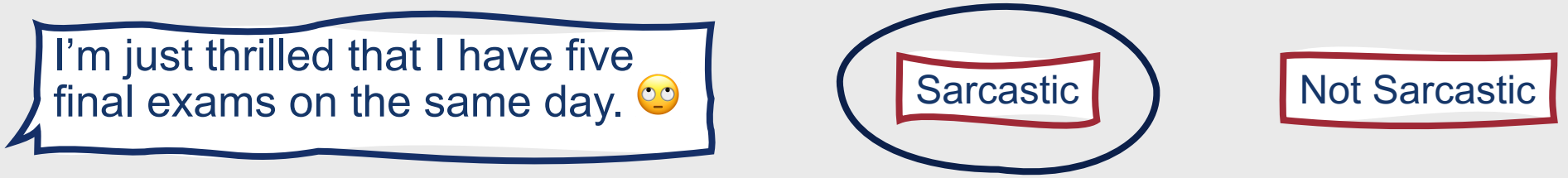
Example: Cross-Entropy Loss



Instance	Predicted Probability: Sarcastic	Predicted Probability: Not Sarcastic	Actual Probability: Sarcastic	Actual Probability: Not Sarcastic
I'm just thrilled that I have five final exams on the same day. 🙄	0.96	0.04	1	0

$$loss(y_i, y_i') = -\sum_{c=1}^{|C|} p_{i,c} \log \widehat{p_{i,c}} = -p_{i,sarcastic} \log \widehat{p_{i,sarcastic}} - p_{i,not\ sarcastic} \log \widehat{p_{i,not\ sarcastic}}$$

Example: Cross-Entropy Loss

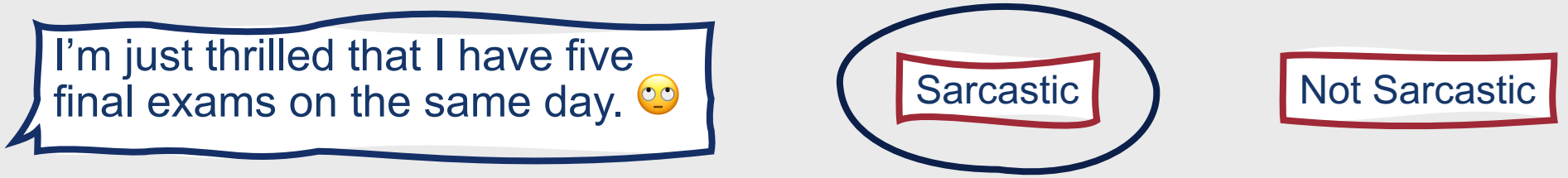


Instance	Predicted Probability: Sarcastic	Predicted Probability: Not Sarcastic	Actual Probability: Sarcastic	Actual Probability: Not Sarcastic
I'm just thrilled that I have five final exams on the same day. 🙄	0.96	0.04	1	0

$$loss(y_i, y_i') = -\sum_{c=1}^{|C|} p_{i,c} \log \widehat{p_{i,c}} = -p_{i,sarcastic} \log \widehat{p_{i,sarcastic}} - p_{i,not\ sarcastic} \log \widehat{p_{i,not\ sarcastic}}$$

$$loss(y_i, y_i') = -1 * \log 0.96 - 0 * \log 0.04$$

Example: Cross-Entropy Loss

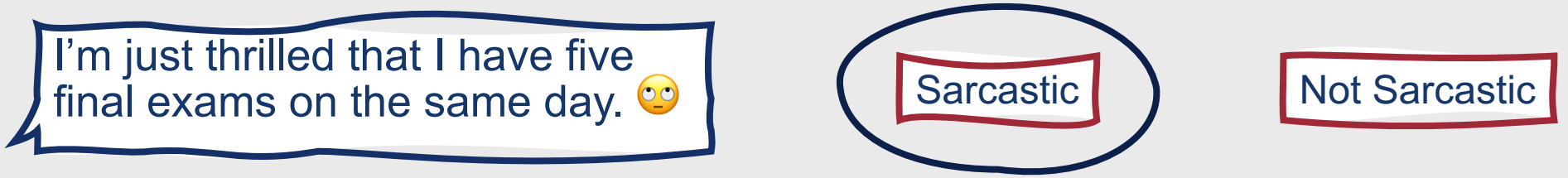


Instance	Predicted Probability: Sarcastic	Predicted Probability: Not Sarcastic	Actual Probability: Sarcastic	Actual Probability: Not Sarcastic
I'm just thrilled that I have five final exams on the same day. 🙄	0.7	0.3	1	0

$$loss(y_i, y_i') = -\sum_{c=1}^{|C|} p_{i,c} \log \widehat{p_{i,c}} = -p_{i,sarcastic} \log \widehat{p_{i,sarcastic}} - p_{i,not\ sarcastic} \log \widehat{p_{i,not\ sarcastic}}$$

$$loss(y_i, y_i') = -1 * \log 0.96 - 0 * \log 0.04 = -\log 0.96 = 0.02$$

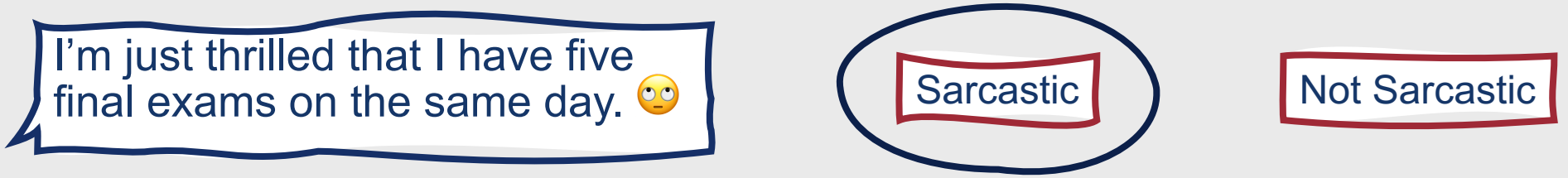
Example: Cross-Entropy Loss



Instance	Predicted Probability: Sarcastic	Predicted Probability: Not Sarcastic	Actual Probability: Sarcastic	Actual Probability: Not Sarcastic
I'm just thrilled that I have five final exams on the same day. 🙄			1	0

What if our predicted values were switched?

Example: Cross-Entropy Loss



Instance	Predicted Probability: Sarcastic	Predicted Probability: Not Sarcastic	Actual Probability: Sarcastic	Actual Probability: Not Sarcastic
I'm just thrilled that I have five final exams on the same day. 🙄	0.04	0.96	1	0

$$loss(y_i, y_i') = -\sum_{c=1}^{|C|} p_{i,c} \log \widehat{p_{i,c}} = -p_{i,sarcastic} \log \widehat{p_{i,sarcastic}} - p_{i,not\ sarcastic} \log \widehat{p_{i,not\ sarcastic}}$$

$$loss(y_i, y_i') = -1 * \log 0.04 - 0 * \log 0.96 = -\log 0.04 = 1.40$$

Greater loss value!