CS 189/289A Discussion 3

Michael Zhang

Updated February 9, 2017

1 Administrative

- Homework 1 grades are out. Homework 2 is due next Monday (go to office hours/HW party if you need help)

2 Review

- High-level view: we've learned a new technique to do classification. Previously, we've talked about techniques that give us decision boundaries (i.e. SVMs, perceptrons). Now, we will take a different approach to classification.
- We model $P(data = X|class = Y)$ for each class Y by specifying a distribution and its parameters for each class Y. To make predictions, we apply Bayes' rule to determine $P(class = Y|data = X)$. This also allows us to incorporate priors and can be more stable than other methods. It also naturally generalizes to the case where we have more than two classes.
- In this framework, we want to make the prediction that minimizes the expected loss.
- A loss function has a natural interpretation—it's the penalty we suffer for making wrong predictions (we generally don't have loss when our prediction is correct). There are situations where certain mistakes are more costly than others (we don't want to say someone is healthy when they have cancer). To account for this, we would set the loss for predicting healthy when a patient has cancer higher. This then leads to a decision boundary that classifies more patients as potentially having cancer (can verify with math).
- To make sure we have notation down (and trying to be consistent with lecture/Problem 1), suppose we have two classes: class 0 and class 1. L(z, y) is the loss when we predict z and the true label is y. If we are predicting class 1 on a data point x, our expected loss is

$$\mathbf{E}[\text{Loss from predicting 1}] = L(1,1) \times P(Y = 1|X = x) + L(1,0) \times P(Y = 0|X = x)$$

  If we assume no loss on correct predictions ($L(1,1) = 0$), this becomes $L(1,0) \times P(Y = 0|X = x)$. Similarly, the expected loss for predicting class 0 is $L(0,1) \times P(Y = 1|X = x)$. What do we predict if $\mathbf{E}[\text{Loss from predicting 1}] < \mathbf{E}[\text{Loss from predicting 0}]$?
- Concept question: do we care about the magnitude of our loss function? Does $L(0,1) = 2$ and $L(1,0) = 5$ lead to the same decisions as $L(0,1) = 20$ and $L(1,0) = 50$? (assuming 0 loss for correct predictions)

3 Problem 1/2

- Logistic function is of the form $f(x) = \frac{1}{1+e^{-x}}$. (Draw out picture) and explain why it has nice properties for representing probability, namely $f(0) = \frac{1}{2}$ and it being a smooth function that approaches 1 when x is large and approaches 0 when x is small. You can think of it as a function that squashes its input into a number between 0 and 1. It will come up again with neural nets.
- Good idea to write down common probability distributions on cheatsheet when studying for exams (such as Gaussian, exponential)

- Problems like these can be messy, but the key takeaway is to set up correctly and carefully work through the algebra.
- We should predict class 1 if:

$$\begin{aligned}\mathbf{E}[\text{Loss from predicting 1}] &< \mathbf{E}[\text{Loss from predicting 0}]\\ L(1,0) \times P(Y=0|X=x) &< L(0,1) \times P(Y=1|X=x)\\ P(Y=0|X=x) &< P(Y=1|X=x)\end{aligned}$$

  Makes sense—if we care equally about both classifying both classes correctly, we predict 1 if the probability of 1 is higher under our models. This is also why 0-1 loss functions are nice to work with.
- Now we actually need to compute $P(Y=1|X=x)$. Using Bayes rule:

$$\begin{aligned}P(Y=1|X=x) &= \frac{P(Y=1 \cap X=x)}{P(X=x)}\\ &= \frac{P(Y=1)P(X=x|Y=1)}{P(Y=1)P(X=x|Y=1)+P(Y=0)P(X=x|Y=0)}\end{aligned}$$

  More generally, remember from probability that the denominator uses Law of Total Probability and sums over the product of the probability of a class and the prior probability we observe x from that class
- Remember that we want something that looks like: $f(x) = \frac{1}{1+e^{-x}}$. What do we do to our expression to make it look more similar?
- Dividing top and bottom by $P(Y=1)P(X=x|Y=1)$ gets us closer to the goal.
- We gotta do some algebra now.

$$\begin{aligned}P(Y=1|X=x) &= \frac{P(Y=1 \cap X=x)}{P(X=x)}\\ &= \frac{1}{1+\frac{P(Y=0)P(X=x|Y=0)}{P(Y=1)P(X=x|Y=1)}}\end{aligned}$$

  It's worthwhile to see that $\frac{P(Y=0)P(X=x|Y=0)}{P(Y=1)P(X=x|Y=1)}$ pops up. What happens to $P(Y=1|X=x)$ when this ratio is greater than 1? Equal to 1? Why does this make sense? Finally,

$$P(Y=1|X=x) = \frac{1}{1+\frac{(1-\pi)\frac{1}{\sqrt{2\pi}\sigma_0}exp(-\frac{(x-\mu_0)^2}{2\sigma_0^2})}{\pi\frac{1}{\sqrt{2\pi}\sigma_1}exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2})}} = \frac{1}{1+\frac{(1-\pi)\sigma_1 exp(-\frac{(x-\mu_0)^2}{2\sigma_0^2})}{\pi\sigma_0 exp(-\frac{(x-\mu_1)^2}{2\sigma_1^2})}}$$

  Unfortunately π is overloaded here but $P(Y=1)=\pi$ and $P(Y=0)=1-\pi$. Note that writing $exp(x)$ instead of e^x can be easier to read on longer expressions.
- To finish, we use $e^{\log x} = x$ on the terms outside of the exponential, and we have a logistic function quadratic in x. It's worth expanding and sanity checking what happens when the σ or the priors are equal. When both are equal, we get the centroid method (good to verify)

4 Problem 3

- In practice, we often need to determine the parameters of our model via training data. Using MLE for Gaussians (shown in lecture) tells us to use the sample mean and sample variance for the mean and variance of our Gaussians.

- Point of confusion in my discussion that I ran out of time to fully explain—Writing out our decision rule: predict class 1 if

$$\mathbf{E}[\text{Loss from predicting 1}] < \mathbf{E}[\text{Loss from predicting 2}]$$

allows us to solve and get a rule for x (predict salmon if weight of new fish is greater than 5.66 pounds). This is the rule that minimizes loss. If we wanted to find the actual loss or probability that we are right, we need to use Bayes' rule:

$$\begin{aligned} P(Y = 1|X = x) &= \frac{P(Y = 1 \cap X = x)}{P(X = x)} \\ &= \frac{P(Y = 1)P(X = x|Y = 1)}{P(Y = 1)P(X = x|Y = 1) + P(Y = 2)P(X = x|Y = 2)} \end{aligned}$$

We have all of these probabilities because we are using a Gaussian model and have estimated the parameters for our model.

- Note that in the problem model we suffer more penalty when we predict salmon and the actual fish is seabass. That's why we predict seabass more in part 4 than in part 3.