1 Risk

1.1 Recap

In lecture, we generalized our notion of loss from basic 0/1 loss to incorporating the idea of risk:

Risk: $R(h) = \mathbb{E}_{XY}[L(Y, h(x)]$

For two class 0/1 loss we then saw our optimal classification function as:

$h^*(x) = \arg\max_{y \in \{0,1\}} P(Y = y|X = x)$

1.2 More General Loss

We will now work through a problem in which we have two-classes, but different loss.

We want to design an automated fishing system that captures fish, classifies them, and sends them off to two different companies, Goldfish, Inc., and Blue Tang, Inc. For some reason we only ever catch goldfish and blue tang. Goldfish, Inc. wants goldfish, and Blue Tang, Inc. wants blue tang. Given only the weights of the fish we catch, we want to figure out what type of fish it is using machine learning! Let us assume that the weight of both goldfish and blue tang are both normally distributed (univariate Gaussian), given by the p.d.f.:

$$P(x|\mu_i, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

We are given this data:

Data for goldfish: $\{3, 4, 5, 6, 7\}$

Data for blue tang: $\{5, 6, 7, 8, 9, 7 + \sqrt{2}, 7 - \sqrt{2}\}$

When we classify blue tang incorrectly, it gets sent to Goldfish, Inc. who won't pay us for the wrong fish and sells it themselves. When we classify goldfish incorrectly, it gets sent to Blue Tang, Inc., who is nice and returns our fish. This situation gives rise to this loss matrix:

	goldfish	blue tang
goldfish	0	100
blue tang	10	0

where the rows represent the predicted, and the columns represent the truth.

1.2.1 MLE

We give you the following maximum likelihood estimates for both classes:

Goldfish(class 1):
$\mu_1 = 5$
$\sigma_1 = \sqrt{2}$

$\pi_1 = \frac{5}{12}$

Blue Tang (class 2):

$\mu_2 = 7$
$\sigma_2 = \sqrt{2}$
$\pi_2 = \frac{7}{12}$

1.2.2 0/1 Loss

Next, find the decision rule when assuming a 0-1 loss function. Recall that a decision rule for the 0-1 loss function will minimize the probability of error.

Since we are under 0/1 loss, we have that decision boundary is where $p(Y = 1|x) = p(Y = 2|x)$.

Thus, $p(Y = 1|x) = \frac{p(x|Y=1)p(Y=1)}{p(x)}$
And, $p(Y = 2|x) = \frac{p(x|Y=2)p(Y=2)}{p(x)}$

So, $\frac{p(x|Y=1)p(Y=1)}{p(x)} = \frac{p(x|Y=2)p(Y=2)}{p(x)}$

And $5p(x|Y = 1) = 7p(x|Y = 2)$

Now, we can use our class conditional densities (and since the sigmas are the same):

$\frac{5}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{2\sigma^2}\right) = \frac{7}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-7)^2}{2\sigma^2}\right)$
Then, we get:
$\ln(5) - \frac{1}{2\sigma^2}(x-5)^2 = \ln(7) - \frac{1}{2\sigma^2}(x-7)^2$

$\ln(5) - \frac{1}{4}(x-5)^2 = \ln(7) - \frac{1}{4}(x-7)^2$

$4\ln(5) - (x-5)^2 = 4\ln(7) - (x-7)^2$
$4\ln(\frac{5}{7}) - x^2 + 10x - 25 = -x^2 + 14x - 49$

$4x = 24 + 4\ln(\frac{5}{7})$

$x = 6 + \ln(\frac{5}{7})$

$x \approx 5.66$

The decision rule is: If $x > 5.66$, classify as Blue Tang! Otherwise classify as Goldfish. Note: Because we had the same variance for both class conditionals, the x^2 term canceled out. If that was not the case, then there would be 3 regions.

1.2.3 Different Loss

Now, find the decision rule using the loss matrix above. Recall that a decision rule, in general, minimizes the risk, or expected loss. Let L_{ij} denote the loss for the i,j entry in the loss matrix.

In the general case, we want to make the decision that minimizes risk. Thus, the decision boundary is located at where the risk of making either decision is equal, or: $R(C_1|x) = R(C_2|x)$

To interpret $R(C_i|x)$, think of this as us observing data x and considering taking action C_i, in this case labelling the data x as class i. If the true class is $Y = j$, where Y denotes the true label by definition, we will incur the loss $L_{i,j}$. Because $P(Y = j|x)$ is the probability that the true class is j, the expected loss associated with taking action C_i is:

$R(C_i|x) = \sum_{j=1}^{C} L_{ij} P(Y = j|x)$

This is saying that the risk of predicting class i is equal to the weighted sum of all classes given our data x multiplied by their probabilities of occurring. The weights are now done by our loss matrix L, a generalization of the 0/1 loss.

In our scenario, this expands to:

$L_{11}P(Y = 1|x) + L_{12}P(Y = 2|x) = L_{21}P(Y = 1|x) + L_{22}P(Y = 2|x)$

This yields:

$100 * P(Y = 2|x) = 10 * P(Y = 1|x)$

As in part b with Bayes' rule, this gives:

$100 * \frac{7}{12}N(7, 2) = 10 * \frac{5}{12}N(5, 2)$

$70\mathcal{N}(7, 2) = 5 * \mathcal{N}(5, 2)$

Solving this like part b), we note that the only change is in the value $\ln(\frac{5}{7})$ which now becomes $\ln(\frac{5}{70})$

So, we get $x = 6 + \ln(\frac{5}{70}) \approx 3.36$. Thus, if the weight is greater than 3.36, we classify it as blue tang and if not, we classify it as goldfish. Note, we have imposed a stricter penalty for classifying a blue tang incorrectly as goldfish. As a result, we will tend to call more fish blue tang, as we are more lenient to getting goldfish wrong as opposed to blue tang.

2 PAC Learning

2.1 Some Important Definitions and Theorems

1. Basic notations:

- **True function** (expert/oracle) $c^* : X \to Y$ (unknown)
- Hypothesis space $\mathcal{H}$ and hypothesis $h \in \mathcal{H} : X \to Y$
- Probability Distribution p^* (unknown)
- Training Dataset $S = x^{(1)}, ...x^{(N)}$

2.True Error (expected risk)

$$R(h) = P_{x \sim p^*(x)}(c^*(x) \neq h(x))$$

3.Train Error (empirical risk)

$$\hat{R}(h) = P_{x \sim S}(c^*(x) \neq h(x)) = \frac{1}{N} \sum_{i=1}^{N} = \mathbb{1}(c^*(x^{(i)}) \neq h(x^{(i)})) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y^{(i)}) \neq h(x^{(i)}))$$

4. **PAC criterion** is that we produce a high accuracy hypothesis with high probability. More formally,

$$P(\forall h \in \mathcal{H}, |R(h) - \hat{R}(h)| \leq \varepsilon) \geq 1 - \delta$$

5. A hypothesis $h \in \mathcal{H}$ is **consistent** with training data S if $\hat{R}(h) = 0$ (zero training error/correctly classify)

6. **Sample Complexity** is the minimum number of training examples N such that PAC criterion is satisfied for a given ε (arbitrarily small error) and δ (with high probability)

7. Agnostic and Realizable:

- **Realizable** means $c^* \in \mathcal{H}$
- **Agnostic** means c^* may or may not be in $\mathcal{H}$

Theorem 1. *Finite $\mathcal{H}$ and realizable case: $N \geq \frac{1}{\varepsilon}[\log|\mathcal{H}| + \log\frac{1}{\delta}]$ labelled examples are sufficient so that with probability $1 - \delta$, **all** $h \in \mathcal{H}$ with $\hat{R}(h) = 0$ have $R(h) \leq \varepsilon$.*

Proof Sketch:

- Assume k bad hypotheses $h_1, h_2, ..., h_k$ with $R(h) > \varepsilon$
- Consider a single bad hypothesis h_i. The probability it is consistent with first N data points is $\leq (1-\varepsilon)^N$ (**Note: I.I.D. assumption important here**)
- Using union bound and since $k \leq |H|$, the probability of at least one bad hypothesis is $\leq |H|(1-\varepsilon)^N$
- Fact: $1 - x \leq e^{-x}$
- The probability of at least one bad hypothesis is $\leq |H|e^{-\varepsilon N}$
- Final Step: Calculate N such that $|H|e^{-\varepsilon N} \leq \delta$

2.2 PAC-Learnable?

Consider the following problem:

Suppose we want to learn any arbitrary boolean function on M variables, i.e. $\mathcal{H} = \{f : \{0,1\}^M \to \{0,1\}\}$.

If $M = 10$, $\varepsilon = 0.01$, $\delta = 0.001$, how many examples suffice according to Theorem 1?

$\frac{1}{\varepsilon} = 100$
$\log(\frac{1}{\delta}) = \log(1000)$
$|\mathcal{H}| = 2^{2^M}$ Why is this? For M variables we have 2^M different combinations of variables. We also have two possible outputs for each of these combinations, giving 2^{2^M}

Thus, we get $N \geq 100[2^{10}\ln(2) + \ln(1000)] \approx 71669$. Note that the number of samples is exponential in M!