Homework 1

36-465/665, Spring 2021

Due at 6 pm on Thursday, 11 February 2021 (Pittsburgh)

In problems 1–3, we are trying to predict a binary variable Y, which for definiteness we can say is either 0 or 1. Our prediction will be $\hat{Y}$, also 0 or 1. The prediction is based on an input variable ("feature") X, and $\Pr\left(Y=1|X=x\right)=p(x)$ is the probability that Y is 1 when $X=x$.

1. **Binary classification with 0-1 loss** In this problem, we incur a loss of 0 if $Y=\hat{Y}$, and a loss of 1 if $Y \neq \hat{Y}$. Say that $\hat{Y}=1$ when $p(x) \geq 0.5$ and $\hat{Y}=0$ when $p(x) < 0.5$.
   a. (5) Show that the probability of mis-classifying is $p(x)$ if $p(x) < 0.5$, and $1-p(x)$ if $p(x) > 0.5$.
   b. (4) Show that the probability of mis-classifying can be written as $\min\left(p(x), 1-p(x)\right)$ for all x.
   c. (3) Explain why we can set $\hat{Y}=0$ or $\hat{Y}=1$ when $p(x)=0.5$ without changing the probability of error.
   d. (5) Show that the probability of mis-classifiying can be written as $\frac{1}{2}-\left|p(x)-\frac{1}{2}\right|$ for all x.
   e. (5) Find an expression for the risk of the optimal classifier.
   f. (5) When will the risk be zero? Explain.
2. **Decision boundaries for binary classification** In this problem, assume that we are trying to predict a binary variable Y, which for definiteness can be either 0 or 1. We have a 2×2 matrix L_{ij} which tells us the loss we incur when we predict j but the reality is i.
   a. (5) Write down an expression for the expected loss of setting $\hat{Y}=0$, as a function of $p(x)$ and the elements of the L matrix. Explain your reasoning.
   b. (5) Write down an expression for the expected loss of setting $\hat{Y}=1$, similarly to (a).
   c. (5) Give a criterion for when it is better to set $\hat{Y}=0$ and when it is better to set $\hat{Y}=1$. You should be able to manipulate your answer to be of the form "Predict $\hat{Y}=1$ when $p(x) >$" some expression involving the elements of the L matrix.
   d. (5) Explain, in words, what your answer in (c) says about when you should be indifferent between saying $\hat{Y}=0$ and $\hat{Y}=1$.
   e. (5) Show that when $L_{00}=L_{11}=0$ and $L_{01}=L_{10}>0$, the boundary from (d) is precisely at $p(x)=1/2$. Explain this in words.
3. **Randomized decisions** We're in the same setting as problem 2, but now, instead of *deterministically* setting $\hat{Y}=0$ or $\hat{Y}=1$, we *randomly* set $\hat{Y}=1$ with probability $q(x)$, which is not necessarily equal to the probability $p(x)$ that $Y=1$.
   a. (5) Find an expression for the expected loss at a given x. It should involve $q(x)$, $p(x)$, and the elements of the L matrix.
   b. (5) Hold x fixed, and find an expression for the value of $q(x)$ that will minimize the expected loss. Your answer should involve $p(x)$ and the elements of the loss matrix.
   c. (5) Does the possibility of a randomized rule lead to any improvements over what we can do with deterministic rules?
4. **Log loss and maximum likelihood** In this problem, f is the true (but usually unknown) probability density function of a random variable X, and g is a PDF we're considering as a possible guess about x. Suppose that if we say the probability density function is g and the actual value we see is x, we include the loss $-\log g(x)$. This is called the "logarithmic loss function" or just the "log loss".
   a. (5) What sorts of predictions will get large losses and which ones will get small losses? Does this seem reasonable?
   b. (4) Explain why the risk of g is $-\int f(x) \log g(x) dx$.
   c. (4) It can be shown that $\int w(t) \log t dt \leq \log\left(\int t w(t) dt\right)$ for *any* probability distribution w. (This

is a special case of a theorem about convex functions called “Jensen’s inequality”; you don’t have to show it.) In fact, $\int w(t) \log t dt = \log \int t w(t) dt$ if, and only if, w puts probability 1 on one particular value of t, and probability 0 on all others. Use this to show that $\int f(x) \log \frac{g(x)}{f(x)} dx \leq 0$. Explain how $g(x)$ must relate to $f(x)$ for the integral to be 0. *Hint*: Probability densities integrate to 1.

d. (5) Using (c), show that $-\int f(x) \log f(x) dx < -\int f(x) \log g(x) dx$ unless $f(x) = g(x)$ everywhere.

e. (5) Suppose we see independent data points $x_1, \ldots x_n$, all drawn from the same distribution. Explain why the likelihood of the pdf g is $\prod_{i=1}^{n} g(x_i)$. How is this related to the log loss of g? How is minimizing the log loss related to maximizing the likelihood? What does (d) tell you about maximum likelihood estimation?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.