DSE 220: Machine learning

Worksheet 6 — Logistic regression

1. We identified *inherent uncertainty* as one reason why it might be difficult to get perfect classifiers, even with a lot of training data. In which of the following situations is there likely to be a significant amount of inherent uncertainty?

   (a) x is a picture of an animal and y is the name of the animal

   (b) x consists of the dating profiles of two people and y is whether they will be interested in each other

   (c) x is a speech recording and y is the transcription of the speech into words

   (d) x is the recording of a new song and y is whether it will be a big hit

2. A logistic regression model given by parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ is fit to a data set of points $x \in \mathbb{R}^d$ with binary labels $y \in \{-1, 1\}$. Write down a precise expression for the set of points x with

   (a) $\Pr(y = 1|x) = 1/2$

   (b) $\Pr(y = 1|x) = 3/4$

   (c) $\Pr(y = 1|x) = 1/4$

3. *Form of the squashing function.* For $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{1, 2\}$, consider a distribution over $\mathcal{X} \times \mathcal{Y}$ of the following form:

   - $\Pr(y = 1) = \Pr(y = 2) = 1/2$
   - The distribution of x given $y = 1$ is a spherical Gaussian $N(\mu_1, \sigma^2 I_d)$ and the distribution of x given $y = 2$ is $N(\mu_2, \sigma^2 I_d)$. Recall that the density of $N(\mu, \sigma^2 I_d)$ is given by

   $$p(x) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left(-\frac{\|x - \mu\|^2}{2\sigma^2}\right).$$

   Derive a closed-form formula for $\Pr(y = 1|x)$. How does it relate to the squashing function?

4. When using a logistic regression model with two labels, define the *margin* on a point x to be how far its conditional probability is from $1/2$:

   $$\text{margin}(x) = \left|\Pr(y = 1|x) - \frac{1}{2}\right|.$$

   This is a number in the range $[0, 1/2]$.

   For any $m \in [0, 1/2]$, define the following two quantities based on a **test set**:

   - $f(m)$: the fraction of test points that have margin $\geq m$
   - $e(m)$: the error rate on test points with margin $\geq m$

   As m grows, how will $f(m)$ and $e(m)$ behave? Would we expect them to increase/decrease? Will they necessarily increase/decrease?