Classification — Generative & Discriminative

Assigned reading: 5.1, 5.2.1, 5.2.2, 5.2.4, 5.3

September 17, 2024

classification

- In contrast to the regression problem, the output is not a real number, but a label:

$$\mathcal{X} \to \mathcal{Y} = \{0, \ldots, K-1\}$$

- The labels can be binary, e.g.

$$\mathbb{R}_+ = \{\text{cholesterol levels}\} \to \{0, 1\}$$

$$\{\text{proteins}\} \times \{\text{proteins}\} \to \{0, 1\}$$

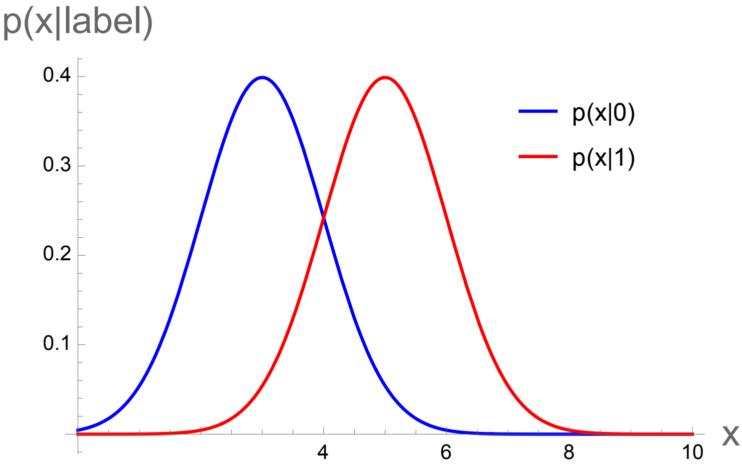
- The labels may not be binary, e.g. the MNIST handwritten digit classification:

$$\{\text{images}\} \to \{0, 1, \ldots, 9\}$$

- Given a dataset $\{(x_i, y_i)\}_{i=1}^n$, we are interested in learning the mapping:

$$f_\theta : \mathcal{X} \to \{0, \ldots, K-1\}$$

class-conditional probabilities



Here the class-conditional probabilities is assumed to be known
(they can be estimated from data).

Bayes Rule

Bayes rule is used to go from class-conditional probabilities to the posterior probabilites

$$p(0|x) = \frac{p(x|0)p(0)}{p(x)},$$

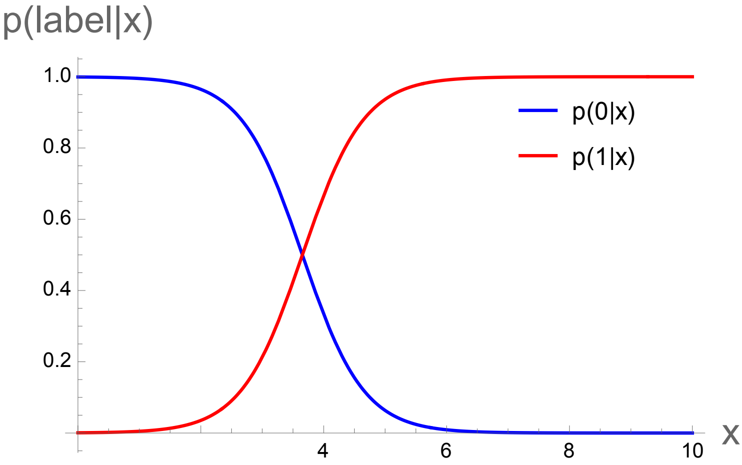
$$p(1|x) = \frac{p(x|1)p(1)}{p(x)},$$

where

$$p(x) = p(x|0)p(0) + p(x|1)p(1).$$

The prior probabilities can also be estimated from data, e.g. $p(0) = 2/3$

posterior probabilities



The **decision boundary** depends on the loss function.

- The rule of thumb that we pick k such that $p(k|x)$ is maximal:

$$k = \operatorname*{argmax}_k p(k|x)$$

is optimal if the goal is to minimize the probability of misclassification:

$$P(error) = \int p(error|x)p(x)dx$$

- To minimize $p(error|x)$ we choose the class with higher posterior probability.
- This is clearly not a good criterion for safety critical problems: the consequences of false negatives are much worse than false positives.
- We will come back to this issue in future lectures.

three ways of building classifiers

- Generative
  - Model the class-conditional probabilities $p(x|k)$.
  - Model the prior $p(k)$.
  - Obtain posteriors $p(k|x)$ using the Bayes rule.
- Discriminative: model $p(k|x)$ directly/explicitly.[1]
- Find decision boundaries $f : \mathcal{X} \to \{0, 1, \ldots, K-1\}$.

[1] This is the inference step. Knowing $p(k|x)$ we can then use Decision Theory to come up with the function $f : \mathcal{X} \to \{0, 1, \ldots, K-1\}$.

logistic regression

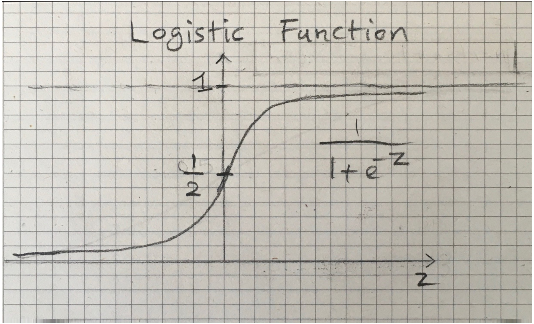
- Logistic regression is a classification method (the output is binary, not a real number)
- In this lecture, we model the posterior probability by a *logistic function* whose argument is a linear function of the features:

$$p(k = 0|x) = \frac{1}{1 + \exp^{-(\theta_1 x + \theta_0)}}$$

- We can justify this choice in multiple ways (and we will derive it for Gaussian class-conditional densities shortly).
- Logistic regression can be extended to K classes.

posterior probability for **Gaussian class-conditional** densities

- IMPORTANT: we assume that variances are the same for different classes
- We do the calculation for 1D feature vectors initially; later we will see that the *d*-dimensional case works out similarly
- We will find that the posterior probability is a logistic (sigmoid) function of $z = \theta_1 x + \theta_0$



☕ (i) Prove that the logistic function is strictly monotonically increasing function of its input and takes values in $(0, 1)$. (ii) prove: $1 - 1/(1 + e^{-z}) = 1/(1 + e^{+z})$

☕☕ Prove that the *posterior* is a *logistic* function

$\boxed{\text{POSTERIOR is LOGISTIC}}$

$$* \quad p(x|k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$$

$$p(0|x) = \frac{p(x|0)\,p(0)}{p(x|0)\,p(0) + p(x|1)\,p(1)}$$

$$= \frac{1}{1 + \frac{p(x|1)}{p(x|0)} \frac{p(1)}{p(0)}}$$

$$= \frac{1}{1 + \exp\left(-\frac{(x-\mu_1)^2 - (x-\mu_0)^2}{2\sigma^2}\right) \frac{p(1)}{p(0)}}$$

$$= \frac{1}{1 + \exp\left(-\left(\left(\frac{\mu_0-\mu_1}{\sigma^2}\right) x - \left(\frac{\mu_1^2-\mu_0^2}{2\sigma^2} - \log\frac{p(1)}{p(0)}\right)\right)\right)}$$

$$= \frac{1}{1+e^{-(\theta_1 x + \theta_0)}}, \text{ where}$$

$$\theta_1 = \frac{\mu_0-\mu_1}{\sigma^2},$$

$$\theta_0 = \frac{\mu_1^2-\mu_0^2}{2\sigma^2} + \log\frac{p(0)}{1-p(0)}$$

- Critical in the proof: the variances σ^2 is the same for $p(x|0)$ and $p(x|1)$
- We do not have to prove this for $p(1|x)$ since

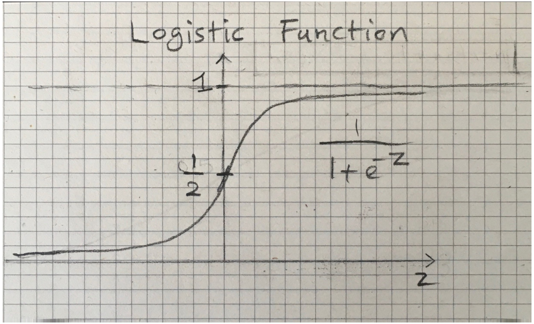
$$\begin{aligned} p(1|x) &= 1 - p(0|x) \\ &= 1 - \frac{1}{1+e^{-z}} \\ &= \frac{1}{1+e^{-(-z)}}, \end{aligned}$$

  where

$$z = \theta_1 x + \theta_2.$$

- Therefore the parameters of the logistic function associated with $p(1|0)$ is obtained by flipping the signs of $\theta = (\theta_0, \theta_1)$ associated with $p(0|1)$.
- Mathematica demonstrations.

graph of the logistic function (as a function of z)



- z is a linear function of x, plus a bias term (referred to as "affine" function):

$$z = \theta_1 x + \theta_0.$$

- Next we generalize this for the multivariate Gaussians, where now $x \in \mathbb{R}^d$, and $z = \theta^\top x + \theta_0$, where we assume

$$X|k \sim \mathcal{N}(\mu_k, \Sigma)$$

Prove that the *posterior* is a *logistic* function of an affine transformation of x.

proof

After inspecting the previous proof (in 1D), we start with the following:

$$\begin{aligned}
\log \frac{p(0|x)}{p(1|x)} &= -\frac{1}{2}(x-\mu_0)^\top \Sigma^{-1}(x-\mu_0) + \frac{1}{2}(x-\mu_1)^\top \Sigma^{-1}(x-\mu_1) + \log \frac{p(0)}{p(1)} \\
&= \underbrace{(\mu_0-\mu_1)^\top \Sigma^{-1}}_{\theta^\top} x + \underbrace{\frac{1}{2}(\mu_1^\top \Sigma \mu_1 - \mu_0^\top \Sigma \mu_0) + \log \frac{p(0)}{p(1)}}_{\theta_0} \\
&= \tilde{\theta}^\top \tilde{x},
\end{aligned}$$

where $\tilde{\theta} = (\theta_0, \theta)$ and $\tilde{x} = (1, x)$. Since $p(1|x) = 1 - p(0|x)$, we have

$$\underbrace{\log \frac{p(0|x)}{1-p(0|x)}}_{\text{log-odds or } \textit{logit}} = \tilde{\theta}^\top \tilde{x} \Rightarrow \frac{1}{p(0|x)} = 1 + exp(-\tilde{\theta}^\top \tilde{x}),$$

Once again, we arrive at the logistic function:[2]

$$p(0|x) = \frac{1}{1 + exp(-\tilde{\theta}^\top \tilde{x})}$$

[2] sanity check: the 1D result from two slides ago is a special case.

a heuristic argument for the logistic

- We like affine functions but can we use them to model posterior probabilities? Linear functions take values in $(-\infty, \infty)$.
  Therefore, the answer is a big NO as we violate the basic rule $0 \leq p \leq 1$.
- Can we do so for odds $p/(1-p)$? Much better since $0 \leq p/(1-p) < \infty$..
- Let's take logarithm to finish the job, i.e. extend the interval to

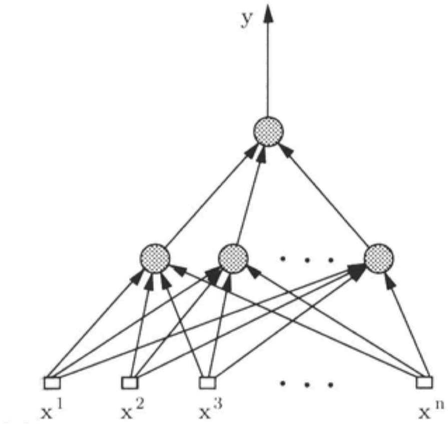
$$\infty < \log \frac{p}{1-p} < \infty$$

- Now we can have a linear model for the log-odds:

$$\text{logit}(p) = \theta^\top x + \theta_0.$$

more in praise of the logistic function

- 1989-1993: Discovery of the universality of function approximation by a sequence of superpositions of logistic sigmoid functions in neural networks.



Modeling the posterior probability distribution

- We say that the class label $Y \in \{0, 1\}$ is Bernoulli random variable, with its probability parameter μ being as above:

$$p(Y = 1|x) = \frac{1}{1 + \exp(-\theta^\top x - \theta_0)} =: \mu(x)$$

- As usual, in the binary case we take y to denote values taken by the random variables:

$$p(y|x) = \mu(x)^y (1 - \mu(x))^{1-y}$$

- Next lecture: we will learn how to estimate (θ, θ_0) with maximum likelihood.