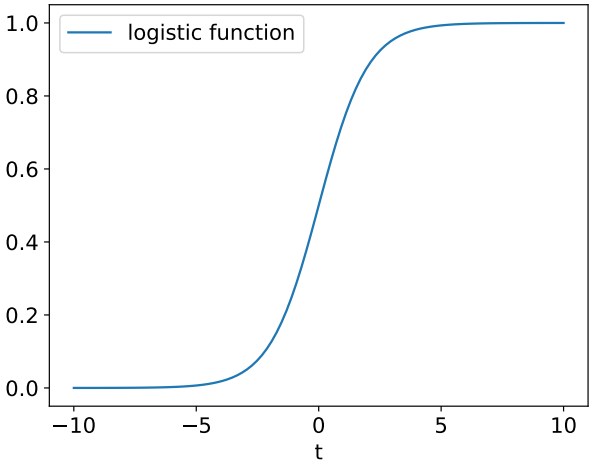
Linear classification

COMS 4771 Fall 2023

Logistic regression model

Logistic regression model: statistical model for binary classification data

- Conditional distribution of Y given $X = x$ is $\text{Bernoulli}(\text{logistic}(w^\mathsf{T} x))$
  - Logistic function: $\text{logistic}(t) = \frac{e^t}{1+e^t}$
  - $1 - \text{logistic}(t) = \frac{1}{1+e^t} = \text{logistic}(-t)$



- Weight vector $w \in \mathbb{R}^d$ is parameter of the model

Log odds ratio (a.k.a. logit) is a linear function of x:

$$\ln\left(\frac{\Pr(Y = 1 \mid X = x)}{\Pr(Y = 0 \mid X = x)}\right) = \underline{\qquad\qquad}$$

Given $X = x$, label 1 is more likely than label 0 if and only if _____

So classifier with smallest error rate (under this distribution) is

$$f^\star(x) = \begin{cases} 1 & \text{if } \underline{\qquad} \\ 0 & \text{if } \underline{\qquad} \end{cases}$$

Such a classifier is called a linear classifier

Often also include an "intercept" parameter b like in linear regression

$$w^\mathsf{T} x + b \quad \text{instead of just} \quad w^\mathsf{T} x$$

- But as usual, we can realize this by including an extra always-1 feature: $x_{d+1} = 1$, and let w_{d+1} act like b

Fitting logistic regression models to data

Maximum likelihood estimation for logistic regression model:

- Likelihood of w:

$$L(w) = \prod_{(x,y)\in\mathcal{S}} \begin{cases} \text{logistic}(w^\mathsf{T}x) & \text{if } y = 1 \\ 1 - \text{logistic}(w^\mathsf{T}x) & \text{if } y = 0 \end{cases}$$
$$= \prod_{(x,y)\in\mathcal{S}} \frac{e^{yw^\mathsf{T}x}}{1 + e^{w^\mathsf{T}x}}$$

- Log-likelihood:

$$\ln L(w) = \sum_{(x,y)\in\mathcal{S}} \left(yw^\mathsf{T}x - \ln(1 + e^{w^\mathsf{T}x}) \right)$$

- Maximizer is not characterized by a system of linear equations, but can be approximated by iterative methods

Iterative improvement algorithm for logistic regression log-likelihood

- Start with $w = 0$
- Repeat T times:

$$w \leftarrow w + \eta \sum_{(x,y)\in\mathcal{S}} (y - \text{logistic}(w^\mathsf{T}x))x$$

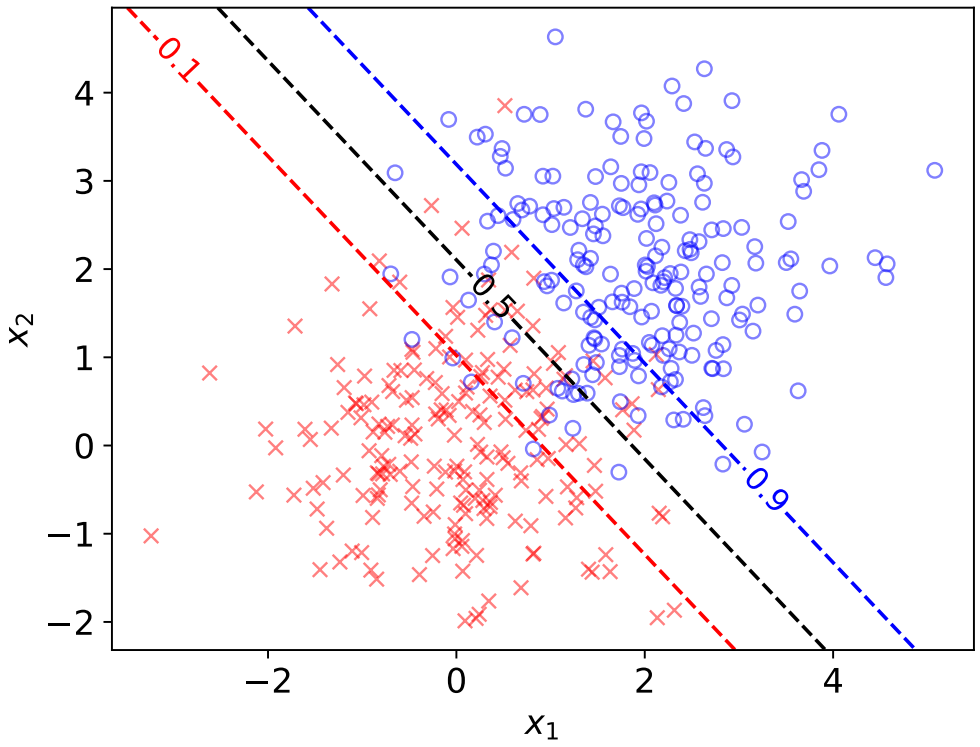
- Return w

(Hyperparameters: "step size" $\eta > 0$, "maximum number of iterations" $T > 0$)

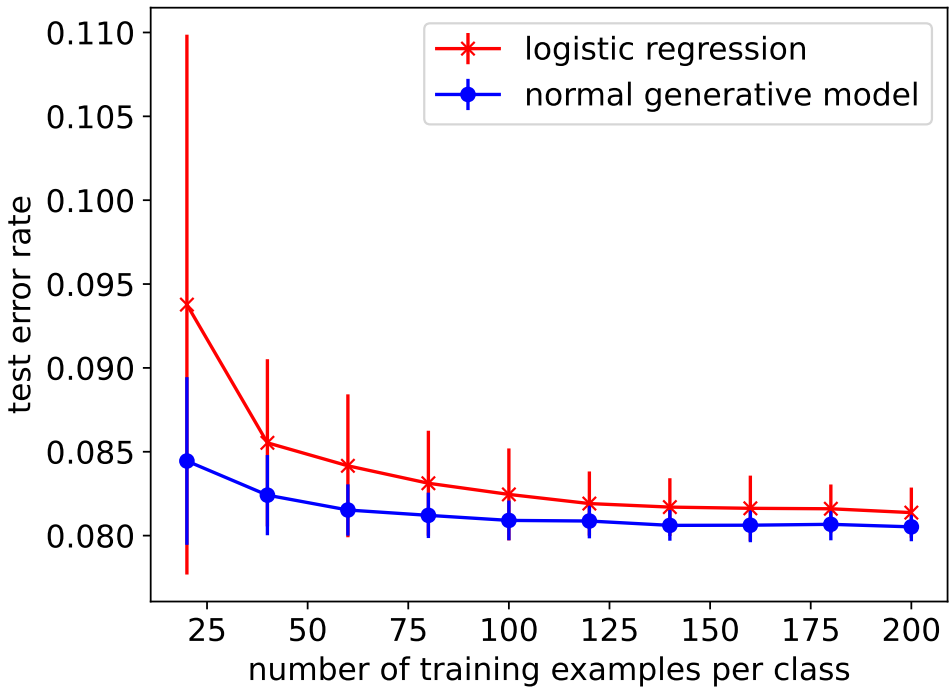
```
def learn(train_x, train_y, eta=0.1, num_steps=1000):
  w = np.zeros(train_x.shape[1])
  for t in range(num_steps):
    w += eta * (train_y - 1/(1+np.exp(-train_x.dot(w)))).dot(train_x)
  return w

def predict(params, test_x):
  return test_x.dot(params) > 0
```


Synthetic example with normal features

- Two classes; class 0: $\mathrm{N}((0,0), I)$, class 1: $\mathrm{N}((2,2), I)$
- 200 training data from each class
- Fit parameters (w_1, w_2, b) of logistic regression model via (approximate) MLE

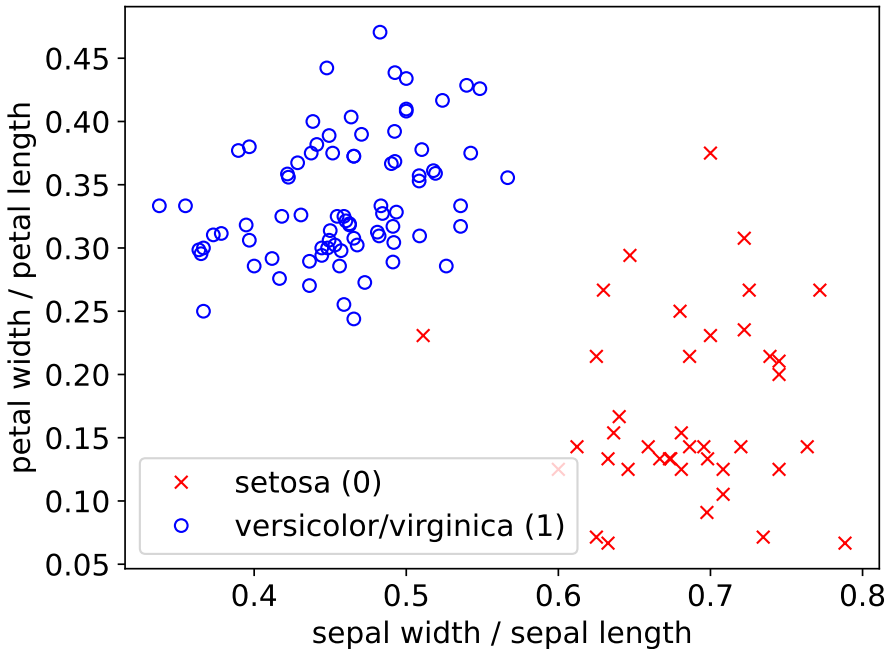


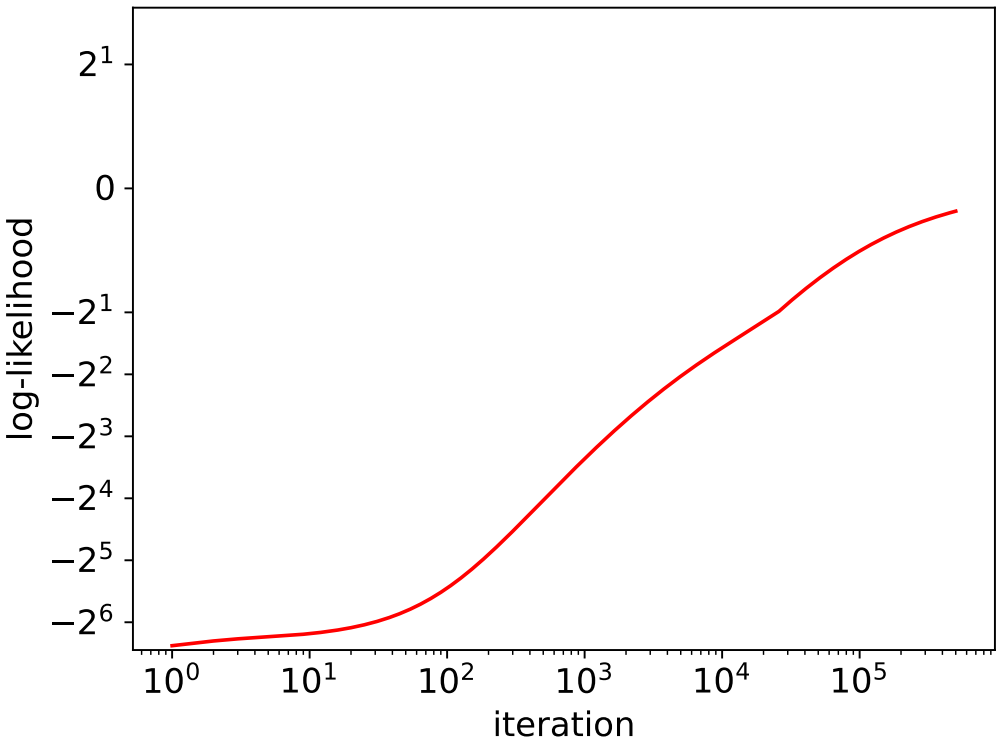
Comparison to normal generative model



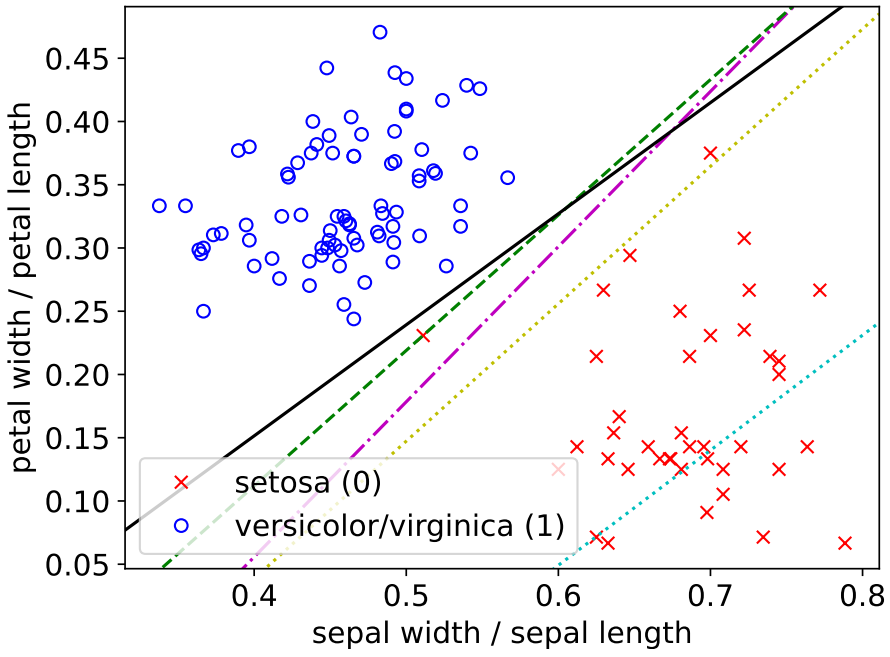
Logistic regression model for iris dataset

Iris dataset, treating versicolor and virginica as a single class

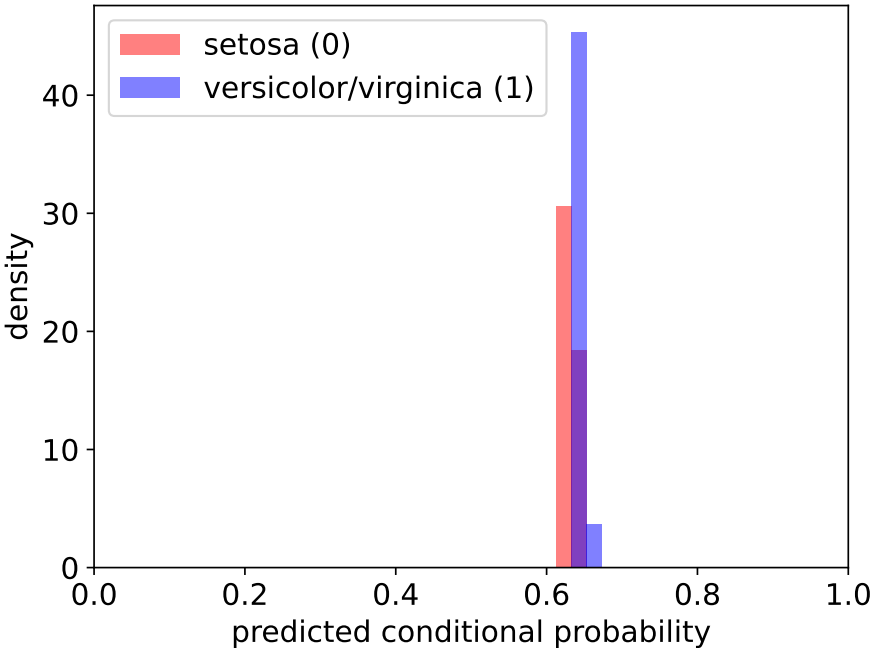




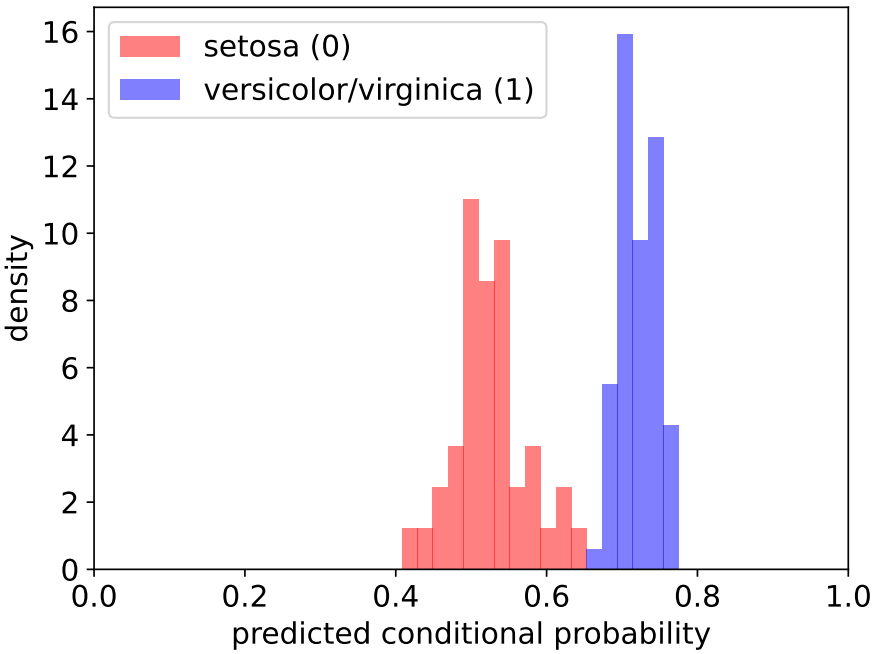
Decision boundary after 50, 500, 5000, 50000, 500000 steps:



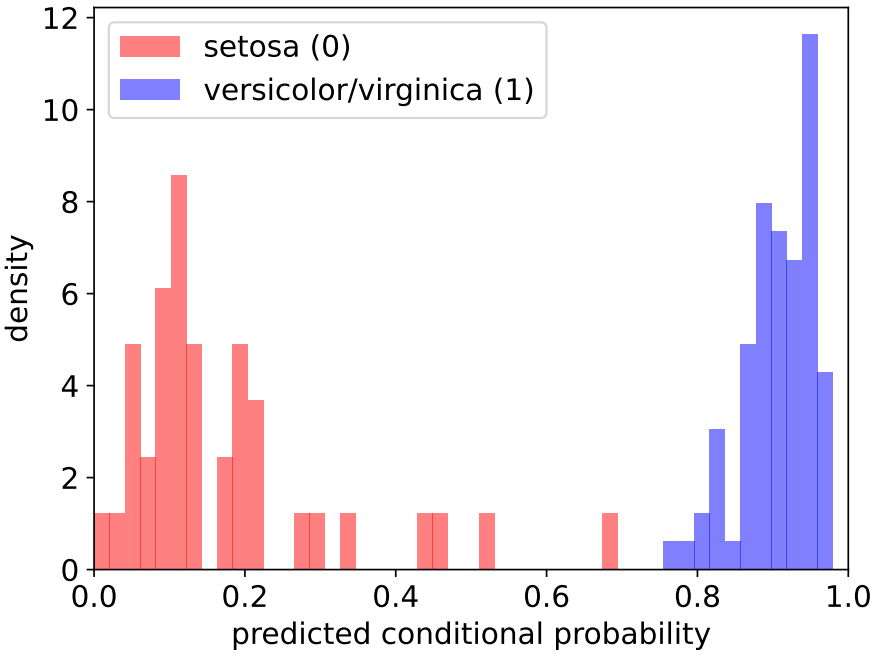
Per-class histograms of predicted conditional probabilities $\widehat{\Pr}(Y = 1 \mid X = x)$ after 5 steps



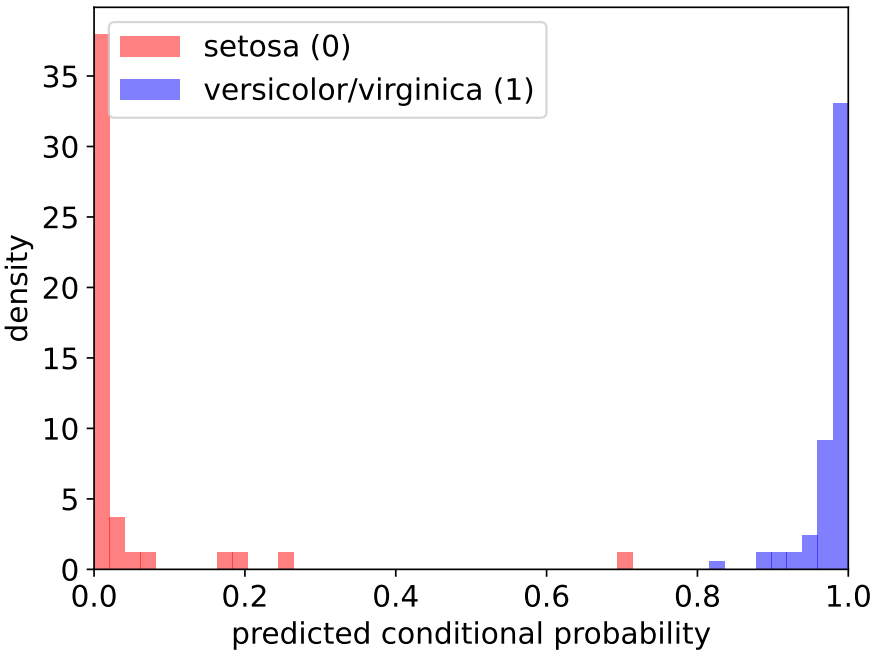
Per-class histograms of predicted conditional probabilities $\widehat{\Pr}(Y = 1 \mid X = x)$ after 50 steps



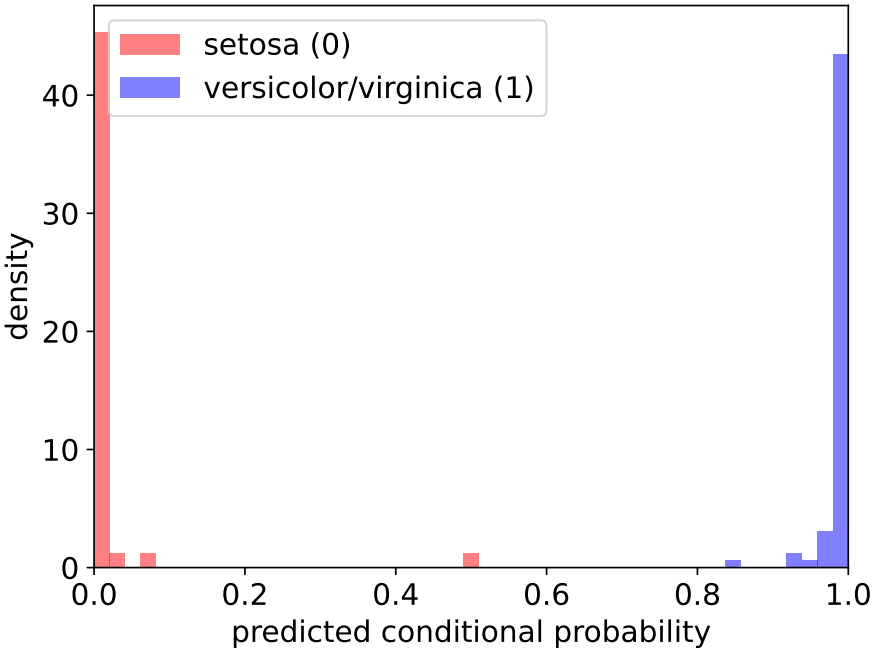
Per-class histograms of predicted conditional probabilities $\widehat{\Pr}(Y = 1 \mid X = x)$ after 500 steps



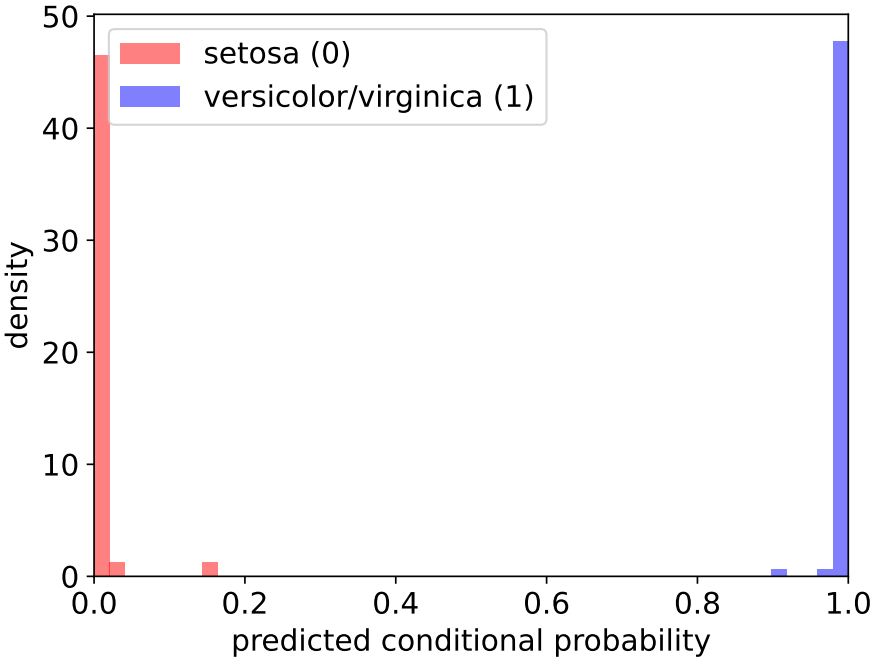
Per-class histograms of predicted conditional probabilities $\widehat{\Pr}(Y = 1 \mid X = x)$ after 5000 steps



Per-class histograms of predicted conditional probabilities $\widehat{\Pr}(Y = 1 \mid X = x)$ after 50000 steps



Per-class histograms of predicted conditional probabilities $\widehat{\Pr}(Y = 1 \mid X = x)$ after 500000 steps



Logarithmic loss

Negative log-likelihood objective in logistic regression can be written as

$$-\ln L(w) = \sum_{(x,y)\in\mathcal{S}} y \ln\left(\frac{1}{\hat{p}_w(x)}\right) + (1-y)\ln\left(\frac{1}{1-\hat{p}_w(x)}\right)$$

where $\hat{p}_w(x) = \text{logistic}(w^\mathsf{T}x)$ is predicted probability of $Y = 1$ given $X = x$ in logistic regression model

Objective is proportional to empirical risk using log(arithmic) loss

$$\text{loss}(\hat{p}, y) = y \ln\left(\frac{1}{\hat{p}}\right) + (1-y)\ln\left(\frac{1}{1-\hat{p}}\right)$$

(a.k.a. cross entropy loss)

Newsgroup dataset

Text classification application: classify articles from internet message boards

- For simplicity, just two classes: religion (class 0), politics (class 1)
- Number of training data: 3028; number of test data: 2017
- Features:
  - Consider a vocabulary of $d = 34250$ "words"
  - Represent an article by vector $x \in \{0,1\}^d$ where

$$x_j = \begin{cases} 1 & \text{if article contains } j\text{-th vocabulary word} \\ 0 & \text{otherwise} \end{cases}$$

  - Log odds ratio in logistic regression model (with $w \in \mathbb{R}^d$):

$$\log \frac{\mathbb{P}(Y=1 \mid x)}{\mathbb{P}(Y=0 \mid x)} = w^T x = \sum_{j=1}^{d} w_j x_j$$

  - Words with negative w_j move log odds ratio to favor class _____
  - Words with positive w_j move log odds ratio to favor class _____

Fit logistic regression model to training data by approximate MLE

- Training error rate: 0.9%; test error rate: 8.5%
- Examining the parameter vector w:

Most negative coefficients		Most positive coefficients	
god	(-1.1291)	israel	$(+0.6019)$
christian	(-0.7304)	gun	$(+0.5766)$
bible	(-0.6747)	government	$(+0.5620)$
jesus	(-0.6679)	american	$(+0.5148)$
keith	(-0.5765)	news	$(+0.4594)$
christians	(-0.5295)	clinton	$(+0.4417)$
religion	(-0.5285)	rights	$(+0.4178)$
church	(-0.4869)	guns	$(+0.4169)$
christ	(-0.4635)	israeli	$(+0.4166)$
athos	(-0.4456)	politics	$(+0.3933)$

Example of article with $\text{logistic}(w^\mathsf{T}x) \approx 0$:

> *Rick, I think we can safely say, 1) Robert is not the only person who understands the Bible, and 2), the leadership of the LDS church historicly never has. Let's consider some "personal interpretations" and see how much trust we should put in "Orthodox Mormonism", which could never be confused with Orthodox Christianity. [...]*

Example of article with $\text{logistic}(w^\mathsf{T}x) \approx 0.5$:

> *Does anyone know where I can access an online copy of the proposed "jobs" or "stimulus" legislation? Please E-mail me directly and if anyone else is interested, I can post this information.*

Example of article with $\text{logistic}(w^\mathsf{T}x) \approx 1$:

```
                    THE ENEMY WITHIN
                    ~~~~~~~~~~~~~~~~

                  By Robert I. Friedman
     The Village Voice, May 11, 1993, Vol. XXXVIII No. 19

   | How The  Anti-Defamation League  Turned the Notion |
   | of Human Rights on Its Head, Spying  on  Progress- |
   | ives and  Funneling Information to Law Enforcement |
   [...]
```

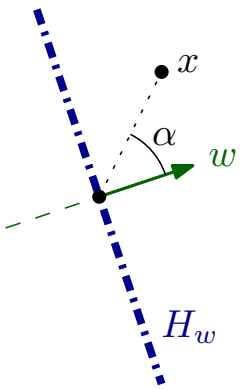
Geometry of linear classifiers

(Homogeneous) linear classifier $f_w \colon \mathbb{R}^d \to \{0, 1\}$, parameterized by $w \in \mathbb{R}^d$:

$$f_w(x) = \mathbb{1}\{w^\mathsf{T} x > 0\} = \begin{cases} 1 & \text{if } w^\mathsf{T} x > 0 \\ 0 & \text{if } w^\mathsf{T} x \leq 0 \end{cases}$$

Decision boundary is a (homogeneous) hyperplane ($d-1$-dimensional subspace):

$$H_w = \{x \in \mathbb{R}^d : w^\mathsf{T} x = 0\}$$

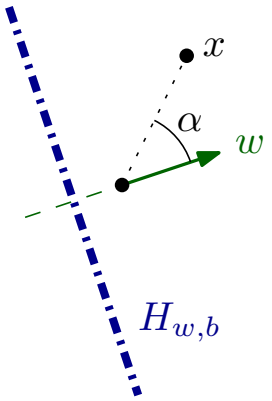


Linear classifier $f_{w,b}\colon \mathbb{R}^d \to \{0,1\}$, parameterized by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$f_{w,b}(x) = \mathbb{1}\{w^\mathsf{T} x + b > 0\} = \begin{cases} 1 & \text{if } w^\mathsf{T} x + b > 0 \\ 0 & \text{if } w^\mathsf{T} x + b \leq 0 \end{cases}$$

Decision boundary is an affine hyperplane:

$$H_{w,b} = \{x \in \mathbb{R}^d : w^\mathsf{T} x + b = 0\}$$



Linear algebra of linear classifiers (assume $w \neq 0$)

- Orthoprojector to $L = \operatorname{span}\{w\}$ is _____
- The (homogeneous) hyperplane H_w is the _____ of L
- The orthogonal projection of x to the line is

  _____

  so _____ is the "coordinate" of x in this line
  (Not technically correct to refer to this "coordinate" as "the projection of x")
- To compute $f_{w,b}(x)$, check if this "coordinate" is more than _____

Computation and Perceptron

Problem: given a dataset $\mathcal{S}$ from $\mathbb{R}^d \times \{0, 1\}$, find $w \in \mathbb{R}^d$ that minimizes training error rate

$$\frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \text{loss}_{0/1}(f_w(x), y)$$

(where $f_w(x) = \mathbb{1}\{w^\mathsf{T} x > 0\}$)

- Computationally intractable in general (assuming P $\neq$ NP)
- Very different from problem solved by OLS

Simpler problem (linear separability): given a dataset $\mathcal{S}$ from $\mathbb{R}^d \times \{0, 1\}$, is there a $w \in \mathbb{R}^d$ such that training error rate of $f_w(x) = \mathbb{1}\{w^\intercal x > 0\}$ is zero?

$$\exists?\ w \in \mathbb{R}^d \text{ s.t. } \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \text{loss}_{0/1}(f_w(x), y) = 0$$

(And return such w if answer is yes)

- "Linear feasibility" problem; can be solved efficiently

Perceptron algorithm

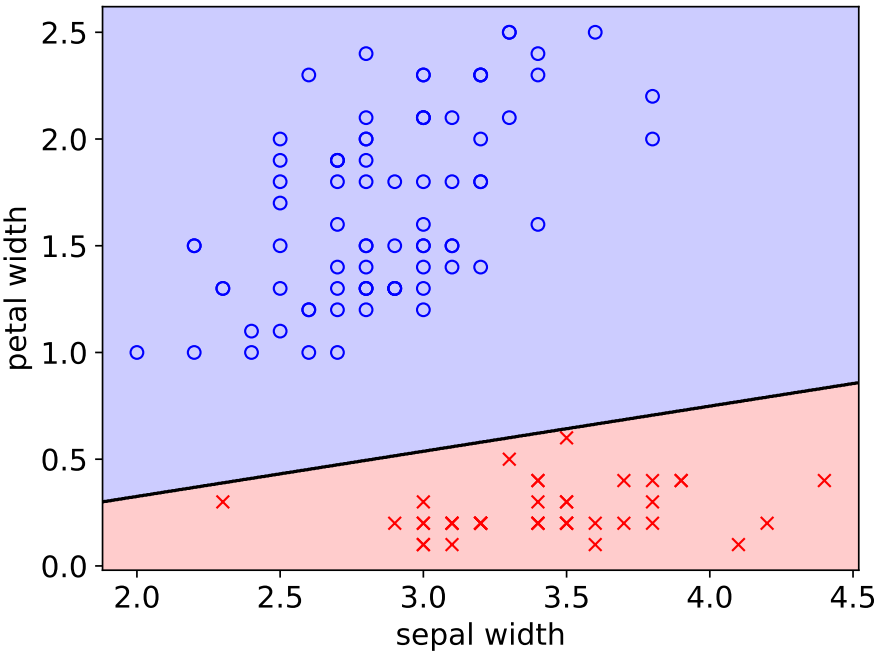
- Start with $w = 0$
- While there exists $(x, y) \in \mathcal{S}$ such that $f_w(x) \neq y$:
  - Let $(x, y) \in \mathcal{S}$ be any such example
  - Update w:

$$w \leftarrow \begin{cases} w + x & \text{if } y = 1 \\ w - x & \text{if } y = 0 \end{cases}$$

- Return w

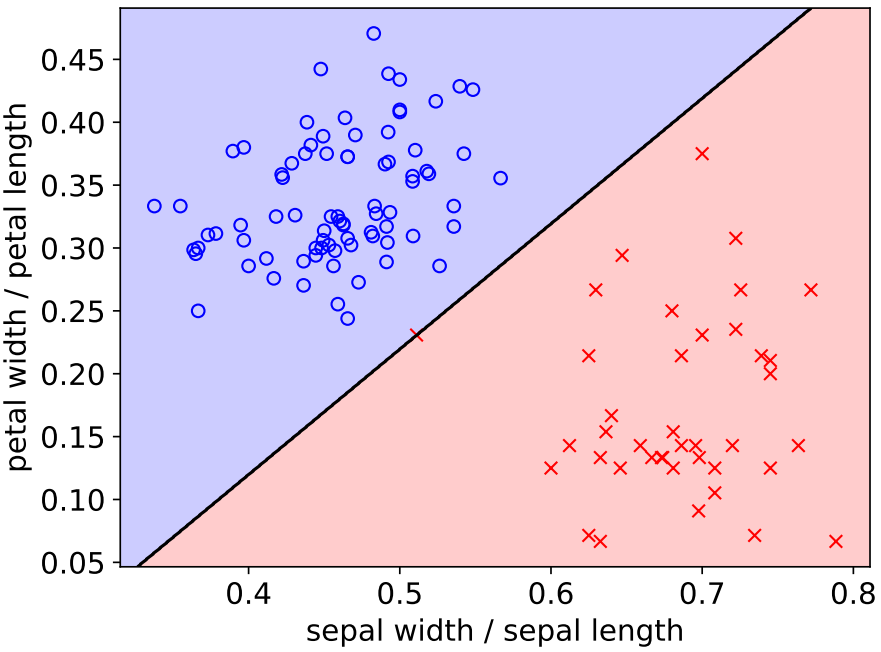
Iris dataset, treating versicolor and virginica as a single class, using features

$x_1 =$ sepal width, $x_2 =$ petal width



Iris dataset, treating versicolor and virginica as a single class, using features

$x_1 =$ sepal width / sepal length, $x_2 =$ petal width / petal length



Convergence guarantee for Perceptron: Suppose there is a unit vector $v \in \mathbb{R}^d$ ($\|v\| = 1$) and a positive number $\gamma > 0$ such that, for all $(x, y) \in \mathcal{S}$,

$$\begin{cases} v^\intercal x \geq +\gamma & \text{if } y = 1 \\ v^\intercal x \leq -\gamma & \text{if } y = 0 \end{cases}$$

Then Perceptron halts after

$$\frac{\max_{(x,y)\in\mathcal{S}} \|x\|^2}{\gamma^2} \text{ loop iterations}$$

(More "wiggle room" $\Longrightarrow$ fewer loop iterations)

Choosing different examples in Perceptron iterations can give different solutions