
CHAPTER 1

A TOUR OF MACHINE LEARNING CLASSIFIERS USING SCIKIT-LEARN

1.1 INTRODUCTION

The goal of this chapter is to use Scikit-Learn to implement some classifiers in order to learn how to use them, their advantages, disadvantages and different use scenarios.

Specifically, we will take a look at 4 popular machine learning models commonly used in academia and in industry. In addition, we will take a look at the Scikit-Learn library, which offers a user-friendly and consistent interface for using those algorithms efficiently and productively.

1.1.1 CHOOSING A CLASSIFICATION ALGORITHM

Each algorithm has its own quirks and relies on certain assumptions. No single classifier works best across all possible scenarios (The Lack of A Priori Distinctions Between Learning Algorithms, Wolpert, David H, Neural Computation 8.7 (1996): 1341-1390).

Observation 1.1.1 (Comparison)

In practice, it is always recommended to compare the behaviour of different algorithms in order to find the best model suitable for a particular problem; these may differ in the number of features or examples, the amount of noise in a dataset, and whether the classes are linearly separable.

Idea 1.1.1

The performance of a classifier relies upon the data that is available for learning. The five main steps that are involved in training a supervised machine learning algorithm can be summarized as follows:

1. Selecting features and collecting labeled training examples
2. Choosing a performance metric
3. Choosing a learning algorithm and training a model
4. Evaluating the performance of the model
5. Changing the settings of the algorithm and tuning the model.

We will mainly focus on the main concepts of the different algorithms in this chapter and revisit topics such as feature selection and preprocessing, performance metrics, and hyperparameter tuning for more detailed discussions later in the book.

1.1.2 FIRST STEPS WITH SCIKIT-LEARN TRAINING A PERCEPTRON

Before we learn about two related learning algorithms: the perceptron and adaline, both implemented in Python using NumPy and other libraries by ourselves.

Now we will take a look at the **scikit-learn API**.

Observation 1.1.2

One of the advantages of using scikit-learn is that combines a user-friendly and consistent interface with a highly optimized implementation of several classification algorithms. Also, this library offers a not only a large variety of learning algorithms, but also many convenient functions to preprocess data and to fine-tune and evaluate our models.

1.1.3 TRAINING A MODEL USING SCIKIT-LEARN

To get started with the scikit-learn library, we will train a perceptron model similar to the one that we implemented in Chapter 2. For simplicity, we will use the already familiar Iris dataset throughout the following sections.

Observation 1.1.3 (Iris Dataset and Its Uses)

Conveniently, the Iris dataset is already available via scikit-learn, since it is a simple yet popular dataset that is frequently used for testing and experimenting with algorithms. Similar to the previous chapter, we will only use two features from the Iris dataset for visualization purposes.

We will assign the petal length and petal width of the 150 flower examples to the feature matrix, X , and the corresponding class labels of the flower species to the vector array, y :

```
1 Class labels: [0 1 2]
2 Class labels: [0 1 2]
```

Code 1.1: Class Labels of Matrix X

The `np.unique(y)` function returned the three unique class labels stored in `iris.target`, and as we can see, the Iris flower class names, `Iris-setosa`, `Iris-versicolor`, and `Iris-virginica`, are already stored as integers (here: 0, 1, 2).

Observation 1.1.4 (Scikit-Learn with Class Labels and String Formats)

Although many scikit-learn functions and class methods also work with *class labels in string format*, using integer labels is a recommended approach to avoid technical glitches and improve computational performance due to a smaller memory footprint; furthermore, encoding class labels as integers is a common convention among most machine learning libraries.

To evaluate how well a trained model performs on unseen data, we will *further split the dataset into separate training and test datasets*.

Idea 1.1.2 (More Info About Best Practices around Model Evaluation)

In Chapter 6, *Learning Best Practices for Model Evaluation and Hyperparameter Tuning*, we will discuss the best practices around model evaluation in more detail.

Using the `train_test_split` function from scikit-learn's `model_selection` module, we randomly split the X and y arrays into 30 percent test data (45 examples) and 70 percent training data (105 examples):

```
1 train_test_split(X, y, test_size=0.3, random_state=1, stratify=y
  )
```

Code 1.2: Train Test Split Function

Observation 1.1.5

Note that the `train_test_split` function already shuffles the training datasets internally before splitting; otherwise, all examples from class 0 and class 1 would have ended up in the training datasets, and the test dataset would consist of 45 examples from class 2. Via the `random_state` parameter, we provided a fixed random seed (`random_state=1`) for the internal pseudo-random number generator that is used for shuffling the datasets prior to splitting. Using such a fixed `random_state` ensures that our results are reproducible.

Lastly, we took advantage of the built-in support for stratification via `stratify=y`. In this context, *stratification means that the `train_test_split` method returns training and test subsets that have the same proportions of class labels as the input dataset.* We can use NumPy's `bincount` function, which counts the number of occurrences of each value in an array, to verify that this is indeed the case:

```
1 >>> print('Labels counts in y:', np.bincount(y))
2 Labels counts in y: [50 50 50]
3 >>> print('Labels counts in y_train:', np.bincount(y_train))
4 Labels counts in y_train: [35 35 35]
5 >>> print('Labels counts in y_test:', np.bincount(y_test))
6 Labels counts in y_test: [15 15 15]
```

Code 1.3: Use of `bincount` Method in Python.

Idea 1.1.3 (Feature Scaling)

Many machine learning and optimization algorithms also require **feature scaling for optimal performance**, as we saw in the gradient descent example in Chapter 2. Here, **we will standardize the features using the `StandardScaler` class from scikit-learn's preprocessing module**:

```
1 sc = StandardScaler()
```

Code 1.4: `StandardScaler` Method in Python.

Using the preceding code, we loaded the `StandardScaler` class from the `preprocessing` module and initialized a new `StandardScaler` object that we assigned to the `sc` variable.

Using the `fit` method, `StandardScaler` estimated the parameters, μ (sample mean) and σ (standard deviation), for each feature dimension from the training data. By calling the `transform` method, we then standardized the training data using those estimated parameters, μ and σ . Note that we used the same scaling parameters to standardize the test dataset so that both the values in the training and test dataset are comparable with one another.

Example 1.1.1

Having standardized the training data, we can now train a perceptron model. Most algorithms in scikit-learn already support multiclass classification by default via the one-versus-rest (OvR) method, which allows us to feed the three flower classes to the perceptron all at once.

Finally, we have the code for the model:

```
1 from sklearn import datasets #To import Iris Dataset
2 import matplotlib.pyplot as plt
```

```

3 from matplotlib.colors import ListedColormap
4 import numpy as np
5 import sklearn.linear_model as linear_models
6 from sklearn.preprocessing import StandardScaler
7 from sklearn.model_selection import train_test_split
8
9 #Plot Regions
10
11 def plot_decision_regions(X, y, classifier, test_idx=None,
12                           resolution=0.2):
13     markers = ('o', 's', '^', 'v', '<')
14     colors = ('red', 'blue', 'lightgreen', 'gray', 'cyan')
15     cmap = ListedColormap(colors[:len(np.unique(y))])
16
17     x1_min, x1_max = X[:, 0].min() - 1, X[:, 0].max() + 1
18     x2_min, x2_max = X[:, 1].min() - 1, X[:, 1].max() + 1
19     xx1, xx2 = np.meshgrid(np.arange(x1_min, x1_max, resolution)
20                             ,
21                             np.arange(x2_min, x2_max, resolution)
22                             )
23     lab = classifier.predict(np.array([xx1.ravel(), xx2.ravel()
24                                       ]).T)
25     lab = lab.reshape(xx1.shape)
26     plt.contourf(xx1, xx2, lab, alpha=0.3, cmap=cmap)
27     plt.xlim(xx1.min(), xx1.max())
28     plt.ylim(xx2.min(), xx2.max())
29
30     for idx, cl in enumerate(np.unique(y)):
31         plt.scatter(x=X[y == cl, 0],
32                     y=X[y == cl, 1],
33                     alpha=0.8,
34                     c=colors[idx],
35                     marker=markers[idx],
36                     label=f'Class {cl}',
37                     edgecolor='black')
38
39     if test_idx is not None:
40         X_test, y_test = X[test_idx, :], y[test_idx]
41         plt.scatter(X_test[:, 0], X_test[:, 1],
42                     c='none', edgecolor='black', alpha=1.0,
43                     linewidth=1, marker='o',
44                     s=100, label='Test set')
45
46 #Import Data
47
48 iris = datasets.load_iris()
49 X = iris.data[:,0:2]
50 y = iris.target
51
52 #Scale Standard Data
53
54 sc = StandardScaler()

```

```

51 X = sc.fit_transform(X)
52
53 #Import Perceptron
54
55 clf = linear_models.Perceptron(eta0=0.1, random_state=1)
56
57 # Split training data and test data
58 indices = np.arange(len(X))
59 X_train, X_test, idx_train, idx_test = train_test_split(
60     X, indices, test_size=0.3, random_state=1
61 )
62 y_train = y[idx_train]
63 y_test = y[idx_test]
64
65 #Fit model with training data
66
67 clf.fit(X_train, y_train)
68
69 #Plot regions
70
71 plot_decision_regions(X, y, clf, idx_test, 0.2)
72 plt.xlabel(iris.feature_names[0] + " (standardized)")
73 plt.ylabel(iris.feature_names[1] + " (standardized)")
74 plt.legend()
75 plt.show()

```

Code 1.5: Perceptron Trained Model using Sklearn

With the slight modification that we made to the `plot_decision_regions` function, we can now specify the indices of the examples that we want to mark on the resulting plots.

As we can see in the resulting plot, the three flower classes can't be perfectly separated by a linear decision boundary:

However, remember from our discussion in Chapter 2 that **the perceptron algorithm never converges on datasets that aren't perfectly linearly separable**, which is *why the use of the perceptron algorithm is typically not recommended in practice*. In the following sections, we will look at more powerful linear classifiers that converge to a loss minimum even if the classes are not perfectly linearly separable.

Observation 1.1.6 (Note)

The Perceptron, as well as other scikit-learn functions and classes, *often has additional parameters that we omit for clarity. You can read more about those parameters using the help function in Python (for instance, `help(Perceptron)`) or by going through the excellent scikit-learn online documentation [here](#).*

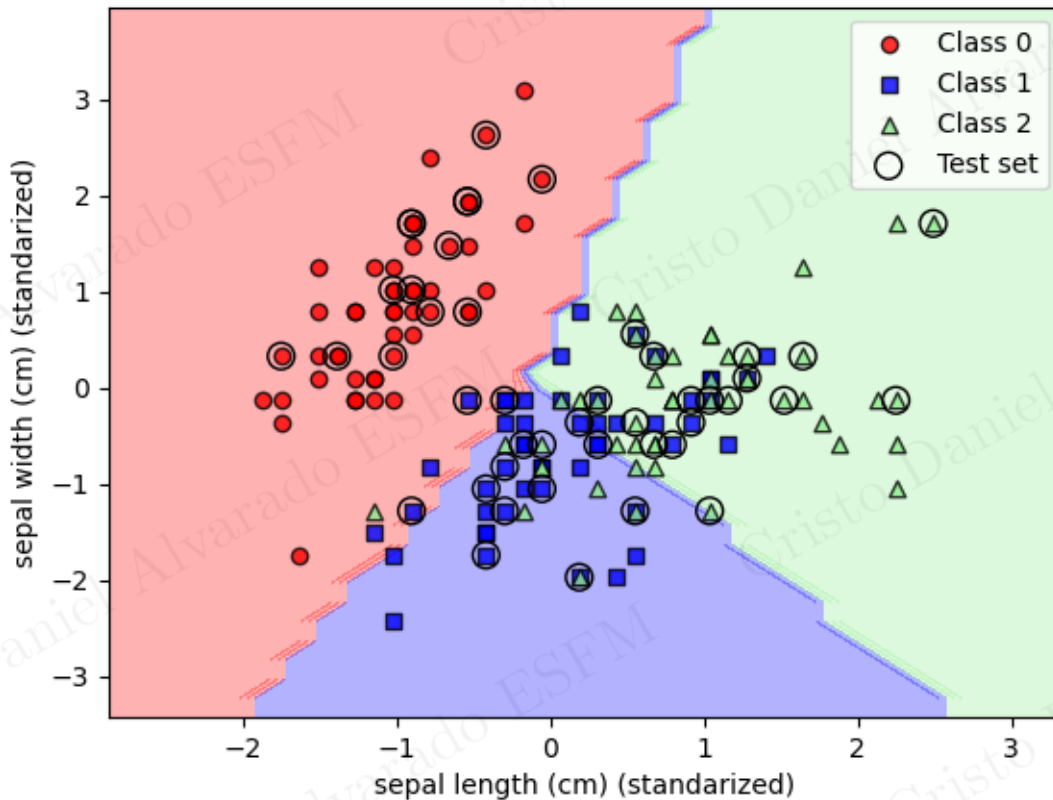


Figure 1.1: Perceptron Trained with Iris Dataset using Scikit-Learn.

1.2 MODELING CLASS PROBABILITIES VIA LOGISTIC REGRESSION

The biggest disadvantage of the Perceptron is that it never converges if the classes aren't linearly separable. The reason for this is that the weights are continuously updated because there is always at least one misclassified training example present in each epoch.

To make better use of our time, we will now look at another simple, yet more powerful, algorithm for linear and binary classification problems: **logistic regression**. Note that, despite its name, *logistic regression is a model for classification, not regression*.

1.2.1 LOGISTIC REGRESSION AND CONDITIONAL PROBABILITIES

Definition 1.2.1 (Logistic Regression)

Logistic regression is a *classification model that is easy to implement and performs very well on linearly separable classes*. It is one of the most widely used algorithms for classification in industry.

Similar to the perceptron and Adaline, *the logistic regression model in this section is also a linear model for binary classification*.

Idea 1.2.1 (Note)

Logistic regression can be generalized to multinomial logistic regression.

To explain the main mechanics behind logistic regression as a probabilistic model for binary classification, let's first introduce the odds: the odds in favor of a particular event. The odds can be

written as $\frac{p}{1-p}$, where p stands for the probability of the positive event. The term "positive event" does not necessarily mean "good," but refers to the event that we want to predict.

Example 1.2.1 (Conditional Probability)

Consider the **probability that a patient has a certain disease given certain symptoms**; we can think of the positive event as class label $y = 1$ and the symptoms as features \mathbf{x} . Hence, for brevity, we can define the probability p as $p := p(y = 1|\mathbf{x})$, the conditional probability that a particular example belongs to class 1 given its features, \mathbf{x} .

We can then further define the logit function, which is simply the logarithm of the odds (log-odds):

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \quad (1.1)$$

Note that \log refers to the natural logarithm, as it is the common convention in computer science. The logit function takes input values in the range 0 to 1 and transforms them into values over the entire real-number range.

Under the logistic model, we assume that there is a linear relationship between the weighted inputs and the log-odds:

$$\text{logit}(p) = w_1x_1 + \cdots + w_nx_n + b = \sum_{i=1}^n w_ix_i + b = \mathbf{w}^T\mathbf{x} + b \quad (1.2)$$

Observation 1.2.1

While the preceding describes an assumption we make about the linear relationship between the log-odds and the net inputs, what **we are actually interested in is the probability p** , the *class-membership probability of an example given its features*. While the logit function maps the probability to a real-number range, *we can consider the inverse of this function to map the real-number range back to a $[0, 1)$ range for the probability p* .

This inverse of the logit function is typically called the **logistic sigmoid function**, which is sometimes simply abbreviated to *sigmoid function* due to its characteristic *S-shape*:

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad \forall z \in \mathbb{R} \quad (1.3)$$

Here, z is the *net input, the linear combination of weights and the inputs* (that is, the features associated with the training examples):

$$z = \mathbf{w}^T\mathbf{x} + b \quad (1.4)$$

Example 1.2.2 (Graph of Sigmoid Function)

Graph of Sigmoid Function:

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 def sigmoid(z):
5     return 1.0 / (1.0 + np.exp(-z))
6
7 z = np.arange(-7, 7, 0.1)
8 sigma_z = sigmoid(z)
9 plt.plot(z, sigma_z)
```

```

10 plt.axvline(0.0, color='k')
11 plt.ylim(-0.1, 1.1)
12 plt.xlabel('z')
13 plt.ylabel('$\sigma(z)$')
14 # y-axis ticks and gridline
15 plt.yticks([0.0, 0.5, 1.0])
16 ax = plt.gca()
17 ax.yaxis.grid(True)
18 plt.tight_layout()
19 plt.show()

```

Code 1.6: Sigmoid Function Graph.

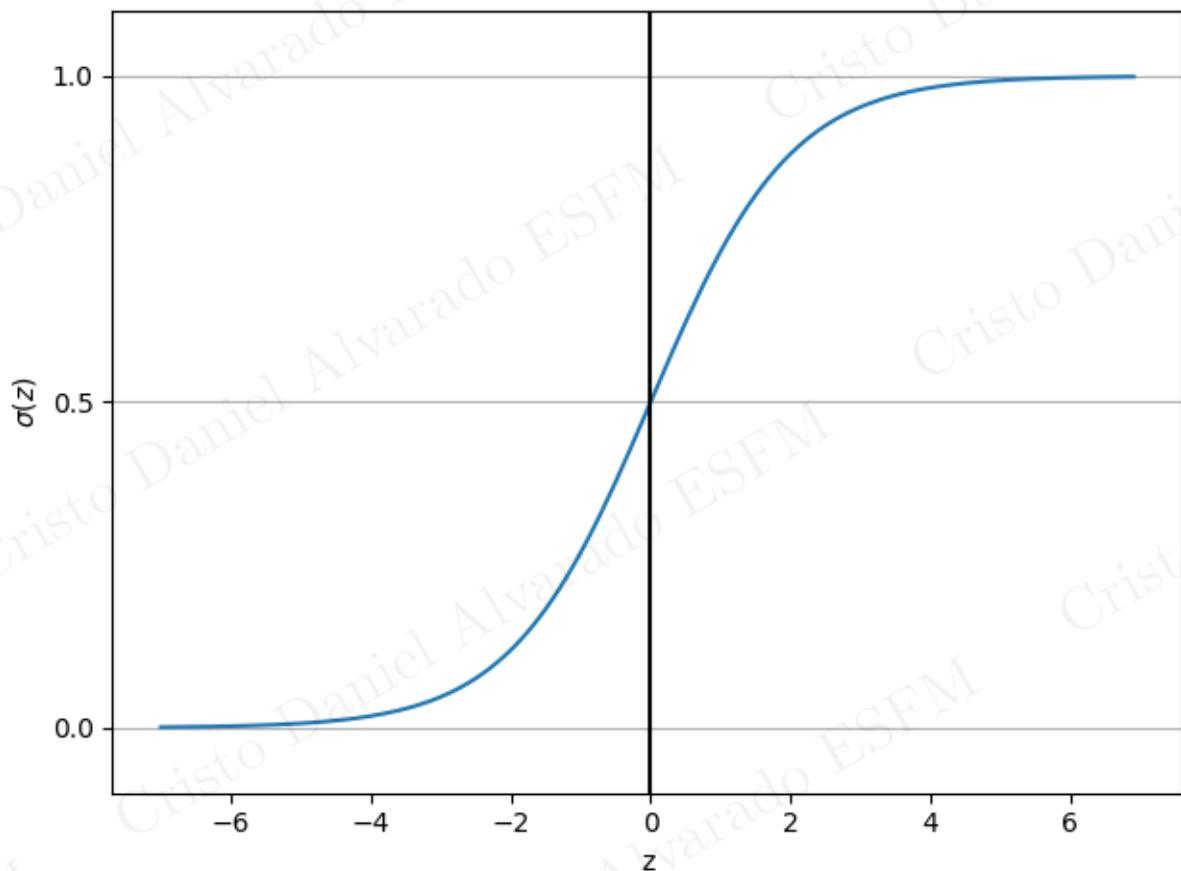


Figure 1.2: Sigmoid Function Graph.

As a result of executing the previous code, you should now see the *S*-shaped (sigmoidal) curve.

Observation 1.2.2 (Comparasion with Adaline)

To build some understanding of the logistic regression model, we can relate it to Adaline. In Adaline, we used the identity function, $\sigma(z) = z$ as the activation function. In logistic regression, *this activation function simply becomes the sigmoid function defined earlier.*

The only difference between Adaline and logistic regression is the activation function.

Definition 1.2.2 (Purpose of an Activation Function)

In simple terms, the purpose of an **activation function** is to decide whether a neuron in a neural network should be "activated" or not, and to what degree. It's what allows neural networks to learn and model complex, non-linear relationships.

The output of the sigmoid function is then interpreted as the probability of a particular example belonging to class 1, $\sigma(z) = p(y = 1|\mathbf{x}; \mathbf{w}, b)$, given its features \mathbf{x} , and parameterized by the weights and bias, \mathbf{w} and b .

For example, if we compute $\sigma(z) = 0.8$ for a particular flower, it means that the chance that this example is an **Iris-versicolor** flower is 80 percent. Therefore, the probability that this flower is an **Iris-setosa** can be calculated as $p(y = 0|\mathbf{x}; \mathbf{w}, b) = 1 - p(y = 1|\mathbf{x}; \mathbf{w}, b) = 0.2$, or 20 percent.

The predicted probability can then be converted into a binary outcome via a threshold function:

$$\hat{y} = \begin{cases} 1 & \text{if } \sigma(z) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

If we look at the preceding plot of the sigmoid function, this is equivalent to the following:

$$\hat{y} = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Observation 1.2.3 (Interest in Predicted Class Labels)

In many applications, we are not only interested in the predicted class labels but also in the class-membership probabilities produced by the sigmoid function before applying the threshold.

Logistic regression is used in weather forecasting, for example, not only to predict whether it will rain on a particular day but also to report the chance of rain. Similarly, logistic regression can be used to predict the chance that a patient has a particular disease given certain symptoms, which is why logistic regression enjoys great popularity in the field of medicine.

1.3 LEARNING THE MODEL WEIGHTS VIA THE LOGISTIC LOSS FUNCTION

You have learned how we can use the logistic regression model to predict probabilities and class labels; now, let's briefly talk about how we fit the parameters of the model, for instance, the weights and bias unit, \mathbf{w} and b . Previously, we defined the mean squared error loss function as follows:

$$L(\mathbf{w}, b|\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \sigma(z^{(i)}))^2$$

We minimized this function in order to learn the parameters for our Adaline classification model. To explain how we can derive the loss function for logistic regression, let's first define the likelihood, \mathcal{L} , that we **want to maximize when we build a logistic regression model**, assuming that the individual examples in our dataset are independent of one another. The formula is as follows:

$$L(w, b | x) = p(y|x; w, b) = \prod_{i=1}^n p(y^{(i)} | x^{(i)}; w, b) = \prod_{i=1}^n (\sigma(z^{(i)}))^{y^{(i)}} (1 - \sigma(z^{(i)}))^{(1-y^{(i)})} \quad (1.5)$$

This is the **likelihood function** $L(w, b | x)$ in terms of a product over n terms. It is expressed as the product of probabilities $p(y^{(i)} | x^{(i)}; w, b)$ for i from 1 to n , which is further expanded to the product of $(\sigma(z^{(i)}))^{y^{(i)}}$ times $(1 - \sigma(z^{(i)}))^{(1-y^{(i)})}$, where σ is the sigmoid function.

Observation 1.3.1

In practice, it is easier to maximize the (natural) log of this equation, which is called the log-likelihood function:

$$l(w, b \mid x) = \log(\mathcal{L}(w, b \mid x)) = \sum_{i=1}^n [y^{(i)} \log(\sigma(z^{(i)})) + (1 - y^{(i)}) \log(1 - \sigma(z^{(i)}))] \quad (1.6)$$

Applying the log function reduces the potential for numerical underflow if the likelihoods are very small. In addition, the product of factors becomes a summation of factors, which makes it easier to obtain the derivative of this function via the addition trick.

1.3.1 DERIVING THE LIKELIHOOD FUNCTION

We can obtain the expression for the likelihood of the model given the data, $L(w, b \mid x)$, as follows. Given that we have a binary classification problem with class labels 0 and 1, we can think of the label 1 as a Bernoulli variable—it can take on two values, 0 and 1, with the probability p of being 1: $Y \sim \text{Bernoulli}(p)$.

For a single data point, we can write this probability as:

$$P(Y = 1 \mid X = x^{(i)}) = \sigma(z^{(i)}) \quad (1.7)$$

$$P(Y = 0 \mid X = x^{(i)}) = 1 - \sigma(z^{(i)}) \quad (1.8)$$

Putting these two expressions together, and using the shorthand $P(Y = y^{(i)} \mid X = x^{(i)}) = p(y^{(i)} \mid x^{(i)})$, we get the probability mass function of the Bernoulli variable: