(d) Forecast next year's prime interest rate.

64 Definition (Probability Distribution)
The vector $\mathbf{p} = \big(p_1, \dots, p_N\big)^\top$ is a probability distribution if:

(i) $0 \le p_k \le 1$ for $k = 1, \dots, N$

(ii) $\sum_{k=1}^{N} p_k = p_1 + \cdots + p_N = 1$

65 Example (One-Hot-Encoding)
Does one-hot-encoding generate probability distributions? Explain?

The softmax function $\mathbf{p} = f_{\text{smax}}(\mathbf{x})$ (defined below) is useful for creating discrete probability distributions from continuous values. The softmax function makes it possible to apply the regression techniques we have developed so far to classification problems. Logistic regression refers to the method of applying a softmax function to the output of linear regression.

66 Definition (Softmax Function)
The **softmax function** $\mathbf{p} = f_{\text{smax}}(\mathbf{x})$ is defined to be

$$f_{\text{smax}}(\mathbf{x}) = \big(\sigma_1(\mathbf{x}) \quad \cdots \quad \sigma_N(\mathbf{x})\big)^\top$$

where

$$\sigma_k(\mathbf{x}) = \frac{e^{x_k}}{e^{x_1} + e^{x_2} + \cdots + e^{x_N}}$$

67 Example (Softmax Function)

(a) Show that for any input $\mathbf{x}$, the output $\mathbf{p} = f_{\text{smax}}(\mathbf{x})$ defines a probability distribution.

(b) Is it possible for $p_k = 0$? $p_k = 1$?

(c) What is $\mathbf{p}$ when $x_1 = x_2 = \cdots = x_N$?

(d) What is $\mathbf{p}$ when one of the x_k is a very large positive number and the other x_k's are very large negative numbers?

(e) Compare the softmax function to one-hot-encoding?

68 Example (Softmax Gives Argmax)
Explain why the softmax function should be called the softargmax function. Hint: Explain how the softmax function "selects" the largest x_k value.

69 Example (Linear Regression)

$$\mathbf{X}_{n\times(d+1)} \to \boxed{\mathbf{w}_{(d+1)\times 1}} \to \hat{\mathbf{y}}_{n\times 1}$$

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$$

70 Example (Logistic Regression)

$$\mathbf{X}_{n\times(d+1)} \to \boxed{\mathbf{W}_{(d+1)\times N}} \to \hat{\mathbf{Y}}_{n\times N} \to \boxed{f_{\text{smax}}} \to \hat{\mathbf{P}}_{n\times N}$$

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\mathbf{W}\\ \hat{\mathbf{P}} &= f_{\text{smax}}(\hat{\mathbf{Y}})\end{aligned}$$

71 Example (Introduction to Keras)
Use `Keras` to predict wine quality of white wine using:
(a) linear regression.
(b) logistic regression.

72 Definition (Jacobian Matrix)
Consider the function

$$\mathbf{y} = f(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_m(x_1, \ldots, x_n) \end{pmatrix}$$

The **Jacobian matrix** of $f(\mathbf{x})$ is defined to be

$$\frac{d\mathbf{y}}{d\mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

73 Example (Jacobian matrix)
Compute the Jacobian matrix $\frac{d\mathbf{y}}{d\mathbf{x}}$ of the function

$$\mathbf{y} = f(\mathbf{x}) = \begin{pmatrix} x_1 x_2 \\ x_1^2 + x_2^2 \end{pmatrix}$$

74 Definition (Chain Rule)
Consider the functions $y = f(z)$ and $z = g(x)$. Define the function $h = f \circ g$ defined below:

$$y = h(x) = (f \circ g)(x) = f(g(x))$$

The derivative $\frac{dy}{dx}$ is given by the chain rule:

$$\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx}.$$

75 Example (Chain Rule)
Let $y = h(x) = \sin(x^2)$. Compute $\frac{dy}{dx}$ using the chain rule

The Jacobian matrix makes it conveniently possible to apply the chain rule to multi-variable, vector valued functions $\mathbf{y} = f(\mathbf{x})$.

76 Theorem (Chain Rule Using Jacobian Matrices)
Consider the functions $\mathbf{y} = f(\mathbf{z})$ and $\mathbf{z} = g(\mathbf{x})$. Define the function $h = f \circ g$ defined below:

$$\mathbf{y} = h(\mathbf{x}) = (f \circ g)(\mathbf{x}) = f(g(\mathbf{x}))$$

$$\mathbf{x} \to \boxed{\mathsf{g}} \to \mathbf{z} \to \boxed{\mathsf{f}} \to \mathbf{y}$$

The derivative $\frac{d\mathbf{y}}{d\mathbf{x}}$ is given by the matrix product of Jacobian matrices:

$$\frac{d\mathbf{y}}{d\mathbf{x}} = \frac{d\mathbf{y}}{d\mathbf{z}}\frac{d\mathbf{z}}{d\mathbf{x}}.$$

Proof for 2×2 Jacobian matrices:

$$\begin{aligned} \frac{d\mathbf{y}}{d\mathbf{z}}\frac{d\mathbf{z}}{d\mathbf{x}} &= \begin{pmatrix} \frac{\partial y_1}{\partial z_1} & \frac{\partial y_1}{\partial z_2} \\ \frac{\partial y_2}{\partial z_1} & \frac{\partial y_2}{\partial z_2} \end{pmatrix} \begin{pmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial y_1}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y_1}{\partial z_2}\frac{\partial z_2}{\partial x_1} & \frac{\partial y_1}{\partial z_1}\frac{\partial z_1}{\partial x_2} + \frac{\partial y_1}{\partial z_2}\frac{\partial z_2}{\partial x_2} \\ \frac{\partial y_2}{\partial z_1}\frac{\partial z_1}{\partial x_1} + \frac{\partial y_2}{\partial z_2}\frac{\partial z_2}{\partial x_1} & \frac{\partial y_2}{\partial z_1}\frac{\partial z_1}{\partial x_2} + \frac{\partial y_2}{\partial z_2}\frac{\partial z_2}{\partial x_2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{pmatrix} \\ &= \frac{d\mathbf{y}}{d\mathbf{x}} \end{aligned}$$

77 Example (Cardinal Sin)
What is the cardinal sin of machine learning?

Answer: Overfitting.