Terminology

Regression - modeling a continuous function
Classification - modeling a discrete categorical function
Supervised learning - learning a function from \((x_i, y_i)\) pairs
Unsupervised learning - learning structure from a data set
Overfitting - model learns peculiarities of training set
Underfitting - model insufficiently complex
Sigmoid function - a function that looks like an "S"
Logistic function \( \frac{1}{1+e^{-x}} \)
Early stopping - step training when validation error starts to increase
Momentum \( v_{t+1} = \mu v_t + \beta \nabla E(\theta_t) \)
Weight decay - L2 reg. on weights
Convolutional neural networks (CNNs)

→ used for image/vision tasks

Image \((1000 \times 1000)\)

\[
\begin{pmatrix}
1 & 2 & 1 & 0 & 3 \\
4 & 5 & -2 & 0 & 2 \\
1 & 3 & 4 & 5 & 3
\end{pmatrix}
\]

Filter \((2 \times 2)\)

\[
\begin{pmatrix}
1 & -1 \\
1 & -1 \\
2 & 2 \\
-2 & -2
\end{pmatrix}
\]

Output

\[
5 & 7 & -1
\]
Activation functions

\[ \sigma(x) = \frac{1}{1 + e^{-x}} \]

\[ \sigma(x) = \tanh(x) \]
REctified Linear Units (RELU)

\[ f_{\text{elu}}(x) = \max(0, x) \]
Overfitting & Regularization

\[ f(x) = 1000 + 2594x - 922x^2 - 250x^3 \]

\[ L = \sum_j \left( \left| y_j - \hat{y}_j \right| \right)^2 + \lambda \sum_{\text{all weights}} \left| w \right|^2 \]

[Diagram showing a graph with labeled axes: validation, train, iterations, loss, predicted output, true output]
Dropout 2012 (a regularization method)

During training, at every iteration, "drop out" (set to 0) each unit with probability $p$

During prediction, multiply weights by $1-p$
Optimization

- Stochastic Gradient Descent
  \[ W_{t+1} = W_t - \eta \nabla L(W_t) \]
  learning rate

- Momentum
  \[ W_{t+1} = W_t + V_{t+1} \]
  \[ V_{t+1} = -\eta \nabla L(W_t) + \varepsilon V_t \]