Today’s Lecture

- Writing code
- Significance testing
- Ethics
Backpropagation

Forward pass

Backward pass
(calculate gradients with chain rule)
Backpropagation

\[ \hat{y} = \sigma(w \cdot x) \]
Backpropagation

\[ \hat{y} = \sigma(w \cdot x) \]

\[ L = (\hat{y} - y)^2 \]
Backpropagation

\[ w \]
\[ \downarrow \]
\[ x \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow \mathcal{L} \]
\[ =w.x \rightarrow =\sigma(h_1) \rightarrow =h_2-y \rightarrow =h_3^2 \]
Backpropagation

\[ x \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow \mathcal{L} \]

- \( x \rightarrow h_1 = w \cdot x \)  
- \( h_2 = \sigma(h_1) \)  
- \( h_3 = h_2 - y \)  
- \( \mathcal{L} = h_3^2 \)

\[
\frac{d\mathcal{L}}{dw_i} = \frac{d\mathcal{L}}{dh_3} \frac{dh_3}{dh_2} \frac{dh_2}{dh_1} \frac{dh_1}{dw_i}
\]
Backpropagation

\[ w \]
\[ \downarrow \]
\[ x \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow y \]
\[ = w \cdot x \]
\[ = \sigma(h_1) \]
\[ = h_2 - y \]
\[ = h_2^2 \]

\[
\frac{d\mathcal{L}}{dw_i} = \frac{d\mathcal{L}}{dh_3} \frac{dh_3}{dh_2} \frac{dh_2}{dh_1} \frac{dh_1}{dw_i}
\]

\[ = 2h_3 \]
Backpropagation

\[ w \]
\[ \downarrow \]
\[ x \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow \mathcal{L} \]
\[ = w . x \quad = \sigma(h_1) \quad = h_2 - y \quad = h_3 \]

\[ \frac{d\mathcal{L}}{dw_i} = \frac{d\mathcal{L}}{dh_3} \frac{dh_3}{dh_2} \frac{dh_2}{dh_1} \frac{dh_1}{dw_i} \]
\[ = 2h_3 \]
Backpropagation

\[
\begin{align*}
\mathbf{w} & \quad \rightarrow \quad \mathbf{h}_1 \quad \rightarrow \quad \mathbf{h}_2 \quad \rightarrow \quad \mathbf{h}_3 \quad \rightarrow \quad \mathbf{L} \\
\mathbf{x} & \quad \rightarrow \quad \mathbf{h}_1 = \mathbf{w} \cdot \mathbf{x} \quad \rightarrow \quad \mathbf{h}_2 = \sigma(\mathbf{h}_1) \quad \rightarrow \quad \mathbf{h}_3 = \mathbf{h}_2 - \mathbf{y} \quad = \mathbf{h}_3^2 \\
\frac{d\mathcal{L}}{dw_i} & = \frac{d\mathcal{L}}{dh_3} \frac{dh_3}{dh_2} \frac{dh_2}{dh_1} \frac{dh_1}{dw_i} \\
& = 2h_3 \mathbf{1} \sigma(\mathbf{h}_1)(1 - \sigma(\mathbf{h}_1))
\end{align*}
\]
Backpropagation

\[ w \]
\[ \downarrow \]
\[ x \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow \mathcal{L} \]
\[ =w \cdot x \rightarrow =\sigma(h_1) \rightarrow =h_2 - y \rightarrow =h^2_3 \]

\[ \frac{d\mathcal{L}}{dw_i} = \frac{d\mathcal{L}}{dh_3} \frac{dh_3}{dh_2} \frac{dh_2}{dh_1} \frac{dh_1}{dw_i} \]

\[ = 2h_3 \cdot 1 \cdot \sigma(h_1)(1 - \sigma(h_1)) \cdot x_i \]
Backpropagation

\[
\begin{align*}
    w & \quad \downarrow \\
    x & \quad \to \quad h_1 \quad \to \quad h_2 \quad \to \quad h_3 \quad \to \quad \mathcal{L} \\
    =w.x & \quad =\sigma(h_1) \quad =h_2-y \quad =h_3^2 \\
\end{align*}
\]

\[
\frac{d\mathcal{L}}{dw_i} = \frac{d\mathcal{L}}{dh_3} \frac{dh_3}{dh_2} \frac{dh_2}{dh_1} \frac{dh_1}{dw_i} \\
= 2h_3 \ 1 \ \sigma(h_1)(1-\sigma(h_1)) \ x_i
\]
Backpropagation

- Need to store computation graph
- Need to store intermediate values
Autograd

- NumPy with automatic differentiation
from autograd import numpy, scipy, grad

def forward(x, w):
    return scipy.special.expit(numpy.dot(x, w))

def loss_fn(x, y, w):
    return (forward(x, w) - y)**2

calculate_w_grad = grad(loss_fn, 2)

w = numpy.random.standard_normal(size=3)
x = numpy.array([.1,.3,.7])
y = 0
w_grad = calculate_w_grad(x, y, w)
from autograd import numpy, scipy, grad

def forward(x, w):
    return scipy.special.expit(numpy.dot(x, w))

def loss_fn(x, y, w):
    return (forward(x, w) - y)**2

calculate_w_grad = grad(loss_fn, 2)

w = numpy.random.standard_normal(size=3)
x = numpy.array([.1,.3,.7])
y = 0

w_grad = calculate_w_grad(x, y, w)
from autograd import numpy, scipy, grad

def forward(x, w):
    return scipy.special.expit(numpy.dot(x, w))

def loss_fn(x, y, w):
    return (forward(x, w) - y)**2

calculate_w_grad = grad(loss_fn, 2)
from autograd import numpy, scipy, grad

def forward(x, w):
    return scipy.special.expit(numpy.dot(x, w))

def loss_fn(x, y, w):
    return (forward(x, w) - y)**2

calculate_w_grad = grad(loss_fn, 2)

w = numpy.random.standard_normal(size=3)
x = numpy.array([.1,.3,.7])
y = 0
w_grad = calculate_w_grad(x, y, w)
TensorFlow

- Automatic differentiation
- Compilation for speed
- Range of architectures
- Range of training algorithms
import tensorflow as tf

w = tf.Variable(tf.random.normal((3,)))
def forward(x):
    return tf.math.sigmoid(tf.math.reduce_sum(w * x))

def loss_fn(x, y):
    return (forward(x) - y)**2

x = tf.constant([.1,.3,.7])
y = tf.constant(0.)
with tf.GradientTape() as g:
    loss = loss_fn(x, y)
w_grad = g.gradient(loss, w)
import tensorflow as tf

w = tf.Variable(tf.random.normal((3,)))

def forward(x):
    return tf.math.sigmoid(tf.math.reduce_sum(w * x))

def loss_fn(x, y):
    return (forward(x) - y)**2

x = tf.constant([.1,.3,.7])
y = tf.constant(0.)

with tf.GradientTape() as g:
    loss = loss_fn(x, y)
    w_grad = g.gradient(loss, w)
import tensorflow as tf

w = tf.Variable(tf.random.normal((3,)))
def forward(x):
    return tf.math.sigmoid(tf.math.reduce_sum(w * x))

def loss_fn(x, y):
    return (forward(x) - y)**2

x = tf.constant([.1,.3,.7])
y = tf.constant(0.)

opt = tf.keras.optimizers.SGD()
opt.minimize(lambda: loss_fn(x,y), var_list=[w])
Summary

- Backpropagation:
  - Store computation graph
  - Store intermediate values

- Software packages:
  - Automatic backpropagation
  - Automatic compilation
  - Pre-defined architectures
  - Pre-defined training algorithms
Neural Network Research

- Emphasis on empirical performance

Datasets re-used many times → Easy to get inflated results
Neural Network Research

- Emphasis on empirical performance
- Large number of architectures, Large number of hyperparameters
- Datasets re-used many times
Neural Network Research

- Emphasis on empirical performance
- Large number of architectures, Large number of hyperparameters
- Datasets re-used many times
  → Easy to get inflated results
Dror et al. (2018) survey of NLP papers:

<table>
<thead>
<tr>
<th></th>
<th>ACL 2017</th>
<th>TACL 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total papers</td>
<td>196</td>
<td>37</td>
</tr>
<tr>
<td>Experimental papers</td>
<td>180</td>
<td>33</td>
</tr>
</tbody>
</table>
Significance Testing

Dror et al. (2018) survey of NLP papers:

<table>
<thead>
<tr>
<th></th>
<th>ACL 2017</th>
<th>TACL 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total papers</td>
<td>196</td>
<td>37</td>
</tr>
<tr>
<td>Experimental papers</td>
<td>180</td>
<td>33</td>
</tr>
<tr>
<td>– reporting significance</td>
<td>63 (35%)</td>
<td>18 (55%)</td>
</tr>
</tbody>
</table>
Dror et al. (2018) survey of NLP papers:

<table>
<thead>
<tr>
<th></th>
<th>ACL 2017</th>
<th>TACL 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total papers</td>
<td>196</td>
<td>37</td>
</tr>
<tr>
<td>Experimental papers</td>
<td>180</td>
<td>33</td>
</tr>
<tr>
<td>– reporting significance</td>
<td>63 (35%)</td>
<td>18 (55%)</td>
</tr>
<tr>
<td>– correctly</td>
<td>36 (20%)</td>
<td>15 (45%)</td>
</tr>
</tbody>
</table>
p-Values

- Probability the result would be at least this extreme, under the null hypothesis
p-Values

- Probability the result would be at least this extreme, under the null hypothesis

  NOT:

- Probability the null hypothesis is true
Statistical Significance Testing

- Decide on a **null hypothesis**
- Decide on a **test statistic**
- Decide on a **threshold**
Statistical Significance Testing

- Decide on a **null hypothesis**
- Decide on a **test statistic**
- Decide on a **threshold**
- **Significance level**: probability of incorrectly rejecting null hypothesis (assuming null hypothesis)
Statistical Significance Testing

- Decide on a **null hypothesis**
- Decide on a **test statistic**
- Decide on a **threshold**

**Significance level**: probability of incorrectly rejecting null hypothesis (assuming null hypothesis)

**Power**: probability of correctly rejecting null hypothesis (assuming alternative hypothesis)
Parametric Tests

- Test statistic follows known distribution (with known parameters)

\[ t = \frac{\bar{x}_D - \bar{x}_D}{s_D / \sqrt{n}} \]

"Student's t-distribution with \( n - 1 \) degrees of freedom"
Parametric Tests

- Test statistic follows known distribution (with known parameters)

- Paired Student’s t-test:
  - Paired samples (test datapoints)
  - Scores normally distributed
  - Null hypothesis: same mean
Parametric Tests

- Test statistic follows known distribution (with known parameters)

- Paired Student’s t-test:
  - Paired samples (test datapoints)
  - Scores normally distributed
  - Null hypothesis: same mean
  - Test statistic: \( t = \frac{\sqrt{n}}{s_D} \bar{X}_D \)
Parametric Tests

- Test statistic follows known distribution (with known parameters)

- Paired Student’s t-test:
  - Paired samples (test datapoints)
  - Scores normally distributed
  - Null hypothesis: same mean
  - Test statistic: $t = \frac{\sqrt{n}}{s_D} \bar{X}_D$
  - “Student’s t-distribution with $n - 1$ degrees of freedom”
Nonparametric Tests

- No assumptions about distribution
Nonparametric Tests

- No assumptions about distribution

- Sign test:
  - Paired samples (test datapoints)
  - System A better or system B better
  - Null hypothesis: equal chance
Nonparametric Tests

- No assumptions about distribution
- Sign test:
  - Paired samples (test datapoints)
  - System A better or system B better
  - Null hypothesis: equal chance
  - Test statistic: $n$
Nonparametric Tests

- No assumptions about distribution

- Sign test:
  - Paired samples (test datapoints)
  - System A better or system B better
  - Null hypothesis: equal chance
  - Test statistic: $n$
  - Binomial distribution
Multiple Tests

- If we test many systems, we expect some will pass
Multiple Tests

- If we test many systems, we expect some will pass

- Bonferroni correction:
  - Replace nominal significance level
  - $\alpha \rightarrow \frac{\alpha}{m}$
Base Rate Fallacy

- Evaluate 1000 systems
  - 900 similar to baseline
  - 100 better than baseline
Base Rate Fallacy

- Evaluate 1000 systems
  - 900 similar to baseline
  - 100 better than baseline

- Perform statistical test
  - Significance level: 5%
  - Power: 80%
Base Rate Fallacy

- Evaluate 1000 systems
  - 900 similar to baseline
  - 100 better than baseline

- Perform statistical test
  - Significance level: 5% → 45 pass
  - Power: 80% → 80 pass

Probability system is better, given it passed the test: 64%
### Base Rate Fallacy

- Evaluate 1000 systems
  - 900 similar to baseline
  - 100 better than baseline

- Perform statistical test
  - Significance level: 5% → 45 pass
  - Power: 80% → 80 pass

- Probability system is better, given it passed the test: 64%
Base Rate Fallacy

- Evaluate 1000 systems
  - 960 similar to baseline
  - 40 better than baseline

- Perform statistical test
  - Significance level: 5% → 48 pass
  - Power: 80% → 32 pass

- Probability system is better, given it passed the test: 40%
Base Rate Fallacy

- Evaluate 1000 systems
  - 1000 similar to baseline
  - 0 better than baseline

- Perform statistical test
  - Significance level: 5% → 50 pass
  - Power: 80% → 0 pass

- Probability system is better, given it passed the test: 0%
Effect Size

- A significant difference may not be a large difference
Effect Size

- A significant difference may not be a large difference

- e.g. a coin toss
  - Coins not perfectly symmetric
  - Probability of heads not exactly 50%
  - Difference so small we don't care
Publication Bias

- Hard to publish negative results...
Publication Bias

- Hard to publish negative results...
- Authors may hide failed experiments
Summary of Significance Testing

- Significance testing is important but underused in deep learning

- Choice of test:
  - Parametric (e.g. paired Student’s t-test)
  - Nonparametric (e.g. sign test)
  - Multiple tests (e.g. Bonferroni correction)

- Be careful:
  - Base rate fallacy
  - Effect size
  - Publication bias
Ethics in Data Science

- Task
- Data
- Model
- Training
Ethics in Data Science

- Task
- Data
- Model
- Training

Most research
Ethics in Data Science

- Task
- Data
- Model
- Training

What if this goes wrong?

Most research
Task: Predict death from pneumonia
Task: Predict death from pneumonia

Pattern in data: asthma reduces risk
Task: Predict death from pneumonia

Pattern in data: asthma reduces risk

Real reason: asthma patients sent to Intensive Care Unit, reducing risk
Task: Predict death from pneumonia

Pattern in data: asthma reduces risk

Real reason: asthma patients sent to Intensive Care Unit, reducing risk

Shallow models (e.g. logistic regression) → can identify and fix such problems
Bias

- Bias (statistics): expected value differs from true value
- Bias (law): unfair or undesirable prejudice
Bias

“Bias is a social issue first, and a technical issue second.”

(Crawford, 2017)
Demographic Bias

- Region
- Social Class
- Gender
- Age
- Ethnicity
Jørgensen et al. (2015)

- Many NLP tools trained on newspaper text (e.g. Penn Treebank)
Many NLP tools trained on newspaper text (e.g. Penn Treebank)

Test POS-taggers on Twitter data, incl. African-American Vernacular English:

<table>
<thead>
<tr>
<th>Group</th>
<th>Stanf.</th>
<th>Gate</th>
<th>Ark</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAVE</td>
<td>.614</td>
<td>.791</td>
<td>.775</td>
</tr>
<tr>
<td>non-AAVE</td>
<td>.745</td>
<td>.833</td>
<td>.779</td>
</tr>
</tbody>
</table>

(significant differences in bold)
The Guardian (2017):
“Computer says no: Irish vet fails oral English test needed to stay in Australia”

Bias in training data vs. bias in decisions
Collecting and analysing personal data requires consent
Privacy

- Collecting and analysing personal data requires consent
- Personal data must be stored securely
Privacy

- Collecting and analysing personal data requires consent
- Personal data must be stored securely
- Anonymising personal data is hard
Nouwens et al. (2020): “our empirical survey of CMPs [cookie banners] illustrates the extent to which illegal practices prevail”
Privacy

- Narayan and Shmatikov (2007), on the Netflix Prize dataset: “Using the Internet Movie Database as background knowledge, we successfully identified known users”
Privacy

- Narayan and Shmatikov (2007), on the Netflix Prize dataset: “Using the Internet Movie Database as background knowledge, we successfully identified known users”

- Four users sued Netflix
Summary of Ethics

- Bias in:
  - Training data
  - Model predictions
  - Real-world decisions

- Personal data
  - Consent to use of data
  - Access to data
What We’ve Covered

- Writing code
  - Backpropagation
  - Software packages

- Statistical Significance
  - Student’s t-test, Sign test
  - Base rate fallacy

- Ethics
  - Social bias
  - Privacy