Roadmap
So Far

1. Housing Prices - Regression
   - Regression Model
   - Assessing Performance
   - Ridge Regression
   - LASSO

2. Sentiment Analysis – Classification
   - Classification Overview
   - Logistic Regression
   - Bias / Fairness
   - Naïve Bayes
   - Decision Trees
   - Ensemble Methods
Training Data → Feature extraction \( x \) → ML model \( h(x) \) → ML algorithm \( f \) → Quality metric \( y \) → \( \hat{y} \)
Decision Trees
Decision Tree

- **Branch/Internal node**: splits into possible values of a feature
- **Leaf node**: final decision (the class value)

```
Root
6 3

credit?

excellent
2 0
Safe

fair
3 1
Safe

poor
1 2
Risky
```
Best threshold?

Infinite possible values of \( t \)

Income < \( t^* \)

Income = \( t^* \)

Income >= \( t^* \)

Income

$10K

$120K

Safe

Risky
Predicting probabilities

Loan status: Safe Risky

Credit?

excellent 9 2
Safe

fair 6 9
Risky

poor 3 1
Safe

\[ P(y = \text{Safe} \mid x) = \frac{3}{3 + 1} = 0.75 \]
Probabilities (Depth 1)

\[
\mathbb{P}(\hat{y}_i = -1) = \frac{13}{16} \quad \mathbb{P}(\hat{y}_i = +1) = \frac{11}{15}
\]
Depth 2
Compare Decision Boundaries
Overfitting

- Deep decision trees are prone to overfitting
  - Decision boundaries are interpretable but not stable
  - Small change in the dataset leads to big difference in the outcome

- Overcoming Overfitting:
  - Early stopping
    - Fixed length depth
    - Stop if error does not considerably decrease
  - Pruning
    - Grow full length trees
    - Prune nodes to balance a complexity penalty
Early Stopping

- **Stopping Rules:**
  - 1) All data in the subset have the same label
  - 2) No more features left to split

- **Early Stopping Rule**
  - Only grow up to a max depth hyperparameter (choose via validation)
  - Don’t split if there is not a sufficient decrease in error
  - **Require a minimum number of examples in a leaf node**
    - Will use this on HW
Decision Tree Overview

- **Super Simple:** Interpretable model that is understandable by people without too much ML experience.
- **Very Efficient:** It actually isn’t too hard to train a tree
- **Depth Matters**
  - Too small, it is too weak to learn the function (high bias)
  - Too tall, it is likely to overfit to the data (high variance)
  - Even by choosing depth appropriately, trees tend to not be the best performing models
Ensemble Methods

Bagging – Random Forest
Training Data \( x \) Feature extraction \( h(x) \) ML model \( \hat{y} \)

\( y \)

ML algorithm \( \hat{f} \)

Quality metric

\( \hat{y} = f(x) \)
Instead of switching to a brand new type of model that is more powerful than trees, what if we instead tried to make the tree into a more powerful model.

What if we could combine many weaker models in such a way to make a more powerful model?

A **model ensemble** is a collection of (generally weak) models that are combined in such a way to create a more powerful model.

There are two common ways this is done with trees
- Random Forest (Bagging)
- AdaBoost (Boosting)
A Random Forest is a collection of $T$ Decision Trees. Each decision tree casts a “vote” for a prediction and the ensemble predicts the majority vote of all of its trees.
If I just have one dataset, how could I learn more than one tree?

Solve this with **bootstrapping**! Can create many similar datasets by randomly sampling *with replacement*.

Technically, you also randomly select features too!
The Random Forest model is a specific type of ensemble model that uses \textbf{bagging} (bootstrapped aggregation).

When training the trees on the bootstrapped samples, we actually want to use very deep trees that overfit!

- That sounds bad at first, but we are trying to take advantage of what it means to have a high variance model (low bias).
- Remember that high variance models have low bias because if you “average out” over all the models you could learn, they will not have bias.
- That is exactly what we are doing here! If we average over a bunch of high variance (overfit) models, to get an ensemble that has low bias and lower variance (if we add more trees)!
Random Forest Algorithm

Training
- Make $T$ random samples of the training data that are the same size as the training data but are sampled with replacement.
- Train a really tall tree on each sampled dataset (overfit).

Predict
- For a given example, ask each tree to predict what it thinks the label should be.
- Take a majority vote over all trees.
Microsoft used Random Forests in their Kinect system to identify the “pose” of a person from the depth camera.
Random Forest Overview

- **Use overfitting to our advantage**: Averaging overfit models can help make a strong model.
- **Versatile**: Works pretty well in a lot of cases and can serve many different purposes.
  - Classification, regression, clustering, feature importance
- **Low Maintenance**: Tends to require less hyper-parameter tuning. Good “out of the box” model.
  - More trees is always better here (but takes longer).
  - Some other hyperparameters, but they tend to have a small affect on performance.
- **Efficient**: Trees can be learned in parallel!
AdaBoost

Boosting
Background

A weak learner is a model that only does slightly better than random guessing.

Kearns and Valiant (1988, 1989):
“Can a set of weak learners create a single strong learner?”

Schapire (1990)
“Yes!”
Overview

AdaBoost is a model similar to Random Forest (an ensemble of decision trees) with two notable differences that impact how we train it quite severely.

- Instead of using high depth trees that will overfit, we limit ourselves to decision stumps.
- Each model in the ensemble gets a weight associated to it, and we take a weighted majority vote

\[
\hat{y} = \hat{f}(x) = \text{sign}\left(\sum_{t=1}^{T} \hat{w}_t \hat{f}_t(x)\right)
\]
Example

Income $>= 1000$

- Yes: SAFE
- No: RISKY

Credit History?

- Bad: RISKY
- Good: SAFE

Savings $>= 100k$

- 3 Years: SAFE
- 5 Years: RISKY

Market Cond.

- Bad: RISKY
- Good: SAFE

\[ \hat{f}_1(x) = +1 \]

\[ \hat{f}_2(x) = -1 \]

\[ \hat{f}_3(x) = -1 \]

\[ \hat{f}_4(x) = +1 \]
With AdaBoost, training is going to look very different.

We train each model in succession, where we use the errors of the previous model to affect how we learn the next one.

To do this, we will need to keep track of two types of weights:

- The first are the $\hat{w}_t$ that we will use as the end result to weight each model.
  - Intuition: An accurate model should have a high weight

- We will also introduce a weight $\alpha_t$ for each example in the dataset that we update each time we train a new model.
  - Intuition: We want to put more weight on examples that seem hard to classify correctly
AdaBoost

Ada Glance

Train

for \( t \) in \([1, 2, ..., T]\):
- Learn \( \hat{f}_t(x) \) based on weights \( \alpha_i \)
- Compute model weight \( \hat{w}_t \)
- Recompute weights \( \alpha_i \)

Predict

\[
\hat{y} = \hat{F}(x) = \text{sign}\left( \sum_{t=1}^{T} \hat{w}_t \hat{f}_t(x) \right)
\]
Start with a dataset and train our first model (a decision stump)
For all the things it gets wrong, increase the weight of that example. For each one that’s right, decrease its weight.
Learning w/ Weighted Data

Before, when we learned decision trees was find the split that minimized classification error.

Now, we want to minimize weighted classification error

\[ WeightedError(f_t) = \frac{\sum_{i=1}^{n} \alpha_i \mathbb{1}\{\hat{f}_t(x_i) \neq y_i\}}{\sum_{i=1}^{n} \alpha_i} \]

If an example \(x_2\) has weight \(\alpha_2 = 3\), this means getting that example wrong is the same as getting 3 examples wrong!

- This will most likely change which split is optimal!
Consider the following weighted dataset, what is the weighted classification error of the optimal decision stump (just one split)?

We want to use the TumorSize and IsSmoker to predict if a patient’s tumor is malignant.

<table>
<thead>
<tr>
<th>TumorSize</th>
<th>IsSmoker</th>
<th>Malignant</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>No</td>
<td>No</td>
<td>0.5</td>
</tr>
<tr>
<td>Small</td>
<td>Yes</td>
<td>Yes</td>
<td>1.2</td>
</tr>
<tr>
<td>Large</td>
<td>No</td>
<td>No</td>
<td>0.3</td>
</tr>
<tr>
<td>Large</td>
<td>Yes</td>
<td>Yes</td>
<td>0.5</td>
</tr>
<tr>
<td>Small</td>
<td>Yes</td>
<td>No</td>
<td>3.3</td>
</tr>
</tbody>
</table>
Consider the following weighted dataset, what is the weighted classification error of the optimal decision stump (just one split)?

We want to use the TumorSize and IsSmoker to predict if a patient’s tumor is malignant.

<table>
<thead>
<tr>
<th>TumorSize</th>
<th>IsSmoker</th>
<th>Malignant</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>No</td>
<td>No</td>
<td>0.5</td>
</tr>
<tr>
<td>Small</td>
<td>Yes</td>
<td>Yes</td>
<td>1.2</td>
</tr>
<tr>
<td>Large</td>
<td>No</td>
<td>No</td>
<td>0.3</td>
</tr>
<tr>
<td>Large</td>
<td>Yes</td>
<td>Yes</td>
<td>0.5</td>
</tr>
<tr>
<td>Small</td>
<td>Yes</td>
<td>No</td>
<td>3.3</td>
</tr>
</tbody>
</table>
### Tumor Size, Is Smoker, Malignant, Weight

<table>
<thead>
<tr>
<th>Tumor Size</th>
<th>Is Smoker</th>
<th>Malignant</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>No</td>
<td>No</td>
<td>0.5</td>
</tr>
<tr>
<td>Small</td>
<td>Yes</td>
<td>Yes</td>
<td>1.2</td>
</tr>
<tr>
<td>Large</td>
<td>No</td>
<td>No</td>
<td>0.3</td>
</tr>
<tr>
<td>Large</td>
<td>Yes</td>
<td>Yes</td>
<td>0.5</td>
</tr>
<tr>
<td>Small</td>
<td>Yes</td>
<td>No</td>
<td>3.3</td>
</tr>
</tbody>
</table>
Real Valued Features

The algorithm is more or less the same, but now we need to account for weights.

Annual Income

$10,000 $200,000

Risky
Safe

Which split is best? Pick the one that maximizes accuracy.
Brain Break
AdaBoost

Ada Glance

Train

for $t$ in $[1, 2, ..., T]$:
- Learn $\hat{f}_t(x)$ based on weights $\alpha_i$
- Compute model weight $\hat{w}_t$
- Recompute weights $\alpha_i$

Predict

$$\hat{y} = \hat{F}(x) = \text{sign} \left( \sum_{t=1}^{T} \hat{w}_t \hat{f}_t(x) \right)$$
Model Weights $\hat{w}_t$

**Goal:** Want to have high weight for models that are very accurate, and low weight for models that are not.

The specific formula used for AdaBoost

$$\hat{w}_t = \frac{1}{2} \ln \left( \frac{1 - \text{WeightedError}(\hat{f}_t)}{\text{WeightedError}(\hat{f}_t)} \right)$$
**Goal:** Increase the weights of data examples that were hard to classify. If we got it wrong, increase the weight, otherwise decrease it.

$$\alpha_i \left\{ \begin{array}{ll} \alpha_i e^{-\hat{w}_t}, & \text{if } \hat{f}_t(x_i) = y_i \\ \alpha_i e^{\hat{w}_t}, & \text{if } \hat{f}_t(x_i) \neq y_i \end{array} \right.$$
AdaBoost
Ada Glance

Train

for $t$ in $[1, 2, ..., T]$:
- Learn $\hat{f}_t(x)$ based on weights $\alpha_i$
- Compute model weight $\hat{w}_t$
- Recompute weights $\alpha_i$

$$\hat{w}_t = \frac{1}{2} \ln \left( \frac{1 - WeightedError(\hat{f}_t)}{WeightedError(\hat{f}_i)} \right)$$

Predict

$$\hat{y} = \hat{F}(x) = \text{sign} \left( \sum_{t=1}^{T} \hat{w}_t \hat{f}_t(x) \right)$$
Normalizing Weights

Generally, the weights for some points get really large/small in magnitude due to how the data is laid out. Numbers in wildly different scales can often cause problems due to finite precision of computers when it comes to real numbers.

Generally we normalize the weights so they sum to 1 to prevent them from getting too small or too big.

\[ \alpha_i \leftarrow \frac{\alpha_i}{\sum_{j=1}^{n} \alpha_j} \]
AdaBoost
Ada Glance

Train

for $t$ in $[1, 2, \ldots, T]$:
- Learn $\hat{f}_t(x)$ based on weights $\alpha_i$
- Compute model weight $\hat{w}_t$
- Recompute weights $\alpha_i$
- Normalize $\alpha_i$

Predict

\[
\hat{y} = \hat{F}(x) = \text{sign} \left( \sum_{t=1}^{T} \hat{w}_t \hat{f}_t(x) \right)
\]
Brain Break
Visualizing AdaBoost
\( t = 1 \)
Learn a Classifier

Start with all data having same weight
Learn a decision stump that minimizes weighted error
- With all the same weights, this is the same as before!

\[
\hat{f}_1(x) = \ldots
\]

Calculate \( \hat{w}_1 \approx 0.61 \)
$t = 1$

Update Data Weights

Compute new weights $\alpha_i$ based on the errors of $\hat{f}_1$

The points with more weight are drawn larger
Now use new weights to learn best stump that minimizes weighted classification error.

\[ \hat{f}_2(x) = \ldots \]

Calculate \( \hat{w}_2 \approx 0.53 \)

Then update weights based on errors.
AdaBoost
Ensemble

If we plot what the predictions would be for each point, we get something that looks like this:
AdaBoost when $t = 30$

Can eventually get 0 training error with a set of weak learners!
This is most likely overfit
AdaBoost

Ada Glance

Train

for $t$ in $[1, 2, ..., T]$:
- Learn $\hat{f}_t(x)$ based on weights $\alpha_i$
- Compute model weight $\hat{w}_t$
- Recompute weights $\alpha_i$
- Normalize $\alpha_i$

Predict

$$\hat{y} = \hat{F}(x) = \text{sign}\left(\sum_{t=1}^{T} \hat{w}_t \hat{f}_t(x)\right)$$
AdaBoost Theorem

Under some technical conditions…

Training error of boosted classifier $\to 0$ as $T \to \infty$

May oscillate a bit

But will generally decrease, & eventually become 0!

Technical condition: The weak learner can do at least slightly better than complete random guessing
Compare

**Decision Tree**

- 39% test error
- 8% training error
- Overfitting

**AdaBoost**

- 32% test error
- 28.5% training error
- Better fit & lower test error
Overfitting?

Boosting tends to be robust to overfitting

But will eventually overfit
Choose $T$?

How do you end up choosing the number of trees $T$ for boosting?

Like always
- Find $T$ that minimizes validation error
- Do cross validation

You can’t
- Find $T$ that minimizes training error
- Find $T$ that minimizes test error
Application

- Boosting, AdaBoost and other variants like gradient boosting, are some of the most successful models to date.
- They are extremely useful in computer vision
  - The standard for face detection
- Used by most winners of ML competitions (Kaggle, KDD Cup, ...)
- Most industry ML systems use a use model ensembles
  - Some with boosting, some with bagging
  - Many times just use 6 different types of models and hand specify their weights.
AdaBoost Overview

- **Powerful!** One of the most powerful set of models for many real world datasets.
  - Typically does better than random forest with the same number of trees.

- **Higher Maintenance:** You do have to tune parameters
  - AdaBoost: Number of trees is technically important, but the model tends to be robust to overfitting in practice.
  - Gradient Boosting: MANY parameters (all important)

- **Expensive:** Boosting is inherently sequential which means its slow to learn ensembles with many trees.
  - Can be made faster with optimized software like XGBoost (UW)
Recap

Theme: Compare two different ways of making ensembles

Ideas:

- Describe what an ensemble model is
- Explain what a random forest is and why adding trees improves accuracy.
- Formalize how AdaBoost combines weighted votes from simple classifiers and how those classifiers are learned.
- Compare/contrast bagging and boosting.
- Describe the steps of the AdaBoost algorithm.