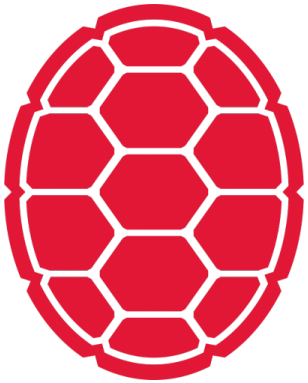


Introduction to Machine Learning

Machine Learning: Jordan Boyd-Graber
University of Maryland
FEATURE ENGINEERING

Content Questions

Announcements

- HW1 Turned in
- HW2 Due on Friday
- HW3 Due in two weeks (up now)

Administrivia Questions

Administrivia Questions

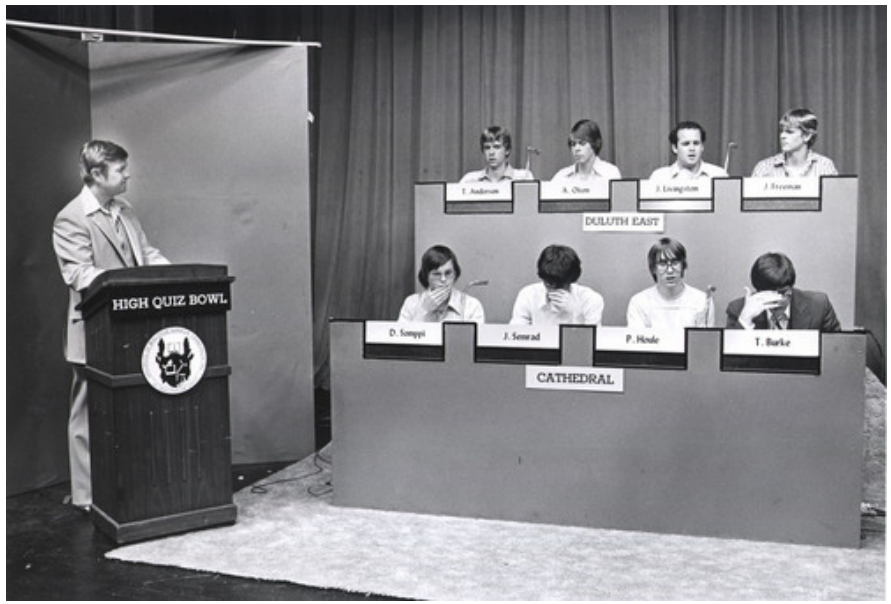
Administrivia Questions

Example of Feature Engineering

- Talk about problem domain: quiz bowl
- Brainstorm features
- Similar to HW3
- Need to get to airport . . .

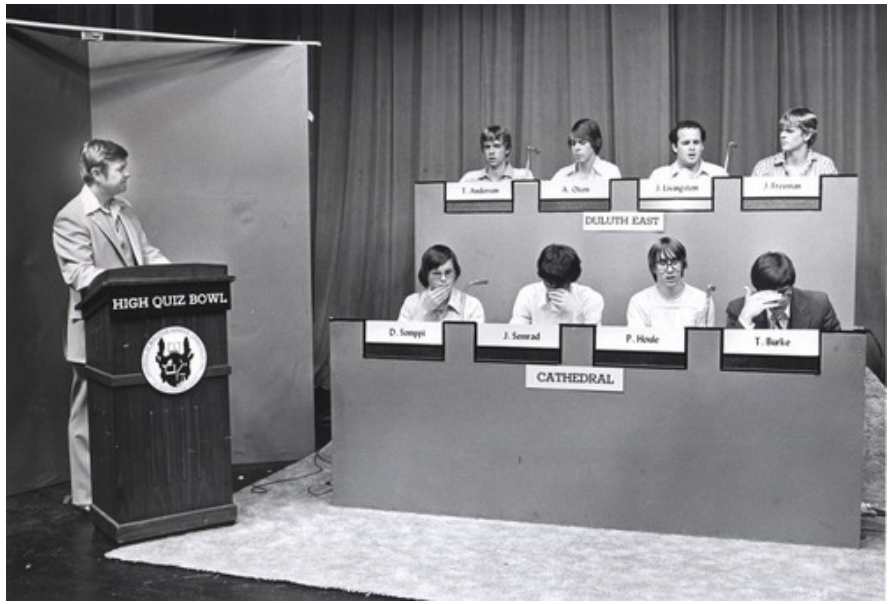
Humans doing Incremental Classification

- Game called “quiz bowl”
- Two teams play each other
  - Moderator reads a question
  - When a team knows the answer, they signal (“buzz” in)
  - If right, they get points; otherwise, rest of the question is read to the other team
- Hundreds of teams in the US alone



Humans doing Incremental Classification

- Game called “quiz bowl”
- Two teams play each other
  - Moderator reads a question
  - When a team knows the answer, they signal (“buzz” in)
  - If right, they get points; otherwise, rest of the question is read to the other team
- Hundreds of teams in the US alone
- Example . . .



Sample Question 1

With Leo Szilard, he invented a doubly-eponymous

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

Sample Question 1

With Leo Szilard, he invented a doubly-eponymous refrigerator with no moving parts. He did not take interaction with neighbors into account when formulating his theory of heat capacity, so Debye adjusted the theory for low temperatures. His summation convention automatically sums repeated indices in tensor products. His name is attached to the A and B coefficients for spontaneous and stimulated emission, the subject of one of his multiple groundbreaking 1905 papers. He further developed the model of statistics sent to him by Bose to describe particles with integer spin. For 10 points, who is this German physicist best known for formulating the special and general theories of relativity?

Albert Einstein

Humans doing Incremental Classification

- This is **not** Jeopardy (Watson)
- There are buzzers, but players can only buzz at the end of a question
- Doesn't discriminate knowledge
- Quiz bowl questions are pyramidal

Feature Engineering

- Determining correct answer is a hard challenge (course project)
- Use same data to do classification
- Divided into specific categories
- Important to know which is which (e.g., might have different submodules to answer each category)

Top Features: Words

```
History          battles:efforts:pyramid:russia:byzantine:pre
Literature       poetic:jew:poet:poem:novel:poems:stories:lit
Social Science   holiday:ritual:zeus:psychological:economist:
Geography        built:nile:hill:square:fault:feature:mountai
Other            code:win:movies:strip:stanley:season:starrin
Biology          selection:enzyme:humans:chromosome:membrane:
Fine Arts        painter:canvas:composers:foreground:commissi
Physics          physical:vector:scattering:classical:quark:p
Chemistry        paradox:doubly:obtained:concentrations:molar
Mathematics      denoted:curve:algebraic:differential:polygon
Earth Science    areas:hot:discontinuity:earth:region:period:
Science          momentum:largest:contrasted:radiation:limit:
Astronomy        found:beyond:object:astronomer:constellation
```

Top Features: Words

```
History          battles:efforts:pyramid:russia:byzantine:pre
Literature       poetic:jew:poet:poem:novel:poems:stories:lit
Social Science   holiday:ritual:zeus:psychological:economist:
Geography        built:nile:hill:square:fault:feature:mountai
Other            code:win:movies:strip:stanley:season:starrin
Biology          selection:enzyme:humans:chromosome:membrane:
Fine Arts        painter:canvas:composers:foreground:commissi
Physics          physical:vector:scattering:classical:quark:p
Chemistry        paradox:doubly:obtained:concentrations:molar
Mathematics      denoted:curve:algebraic:differential:polygon
Earth Science    areas:hot:discontinuity:earth:region:period:
Science          momentum:largest:contrasted:radiation:limit:
Astronomy        found:beyond:object:astronomer:constellation
```

Accuracy: 0.765

Top Features: Bigrams Only

```
History          who served:vice presidential:the election:ch
Literature       his novels:this literary:his plays:poem abou
Social Science   this economic:this anthropologist:of social:
Geography        this state:lake s:the nile:the oldest:was bu
Other            in 2005:this show:an oscar:comic strip:yr am
Biology          amino acid:syndrome and:the enzyme:the brain
Fine Arts        this composer:painter of:opera about:its com
Physics          formula for:a quantum:x rays:of matter:of pa
Chemistry        chemist who:is present:this chemist:tend to:
Mathematics      sub n:data structure:algorithm for:mathemati
Earth Science    its name:mohs hardness:above it:characterize
Science          this process:this number:reaction this:store
Astronomy        the solar:the universe:moon of:the milky:mil
```

Top Features: Bigrams Only

```
History          who served:vice presidential:the election:ch
Literature       his novels:this literary:his plays:poem abou
Social Science   this economic:this anthropologist:of social:
Geography        this state:lake s:the nile:the oldest:was bu
Other            in 2005:this show:an oscar:comic strip:yr am
Biology          amino acid:syndrome and:the enzyme:the brain
Fine Arts        this composer:painter of:opera about:its com
Physics          formula for:a quantum:x rays:of matter:of pa
Chemistry        chemist who:is present:this chemist:tend to:
Mathematics      sub n:data structure:algorithm for:mathemati
Earth Science    its name:mohs hardness:above it:characterize
Science          this process:this number:reaction this:store
Astronomy        the solar:the universe:moon of:the milky:mil
```

Accuracy: 0.800

Top Features: Trigrams Only

History	the battle of:name this leader:ftp what was:
Literature	identify this author:this short story:of the
Social Science	this deity s:name this economic:this thinker
Geography	name this largest:this state contains:this i
Other	on the album:name this film:this comic strip
Biology	disease of the:name this phylum:the producti
Fine Arts	of this painting:this composer of:of this op
Physics	**is** governed by:the standard model:name this
Chemistry	this doubly eponymous:the presence of:of org
Mathematics	name these mathematical:product of two:name
Earth Science	of the paleozoic:d double prime:of this mine
Science	**is** broken down:this compound s:the photoelec
Astronomy	a black hole:**in** the sky:of the universe:brig

Top Features: Trigrams Only

```
History          the battle of:name this leader:ftp what was:
Literature       identify this author:this short story:of the
Social Science   this deity s:name this economic:this thinker
Geography        name this largest:this state contains:this i
Other            on the album:name this film:this comic strip
Biology          disease of the:name this phylum:the producti
Fine Arts        of this painting:this composer of:of this op
Physics          is governed by:the standard model:name this
Chemistry        this doubly eponymous:the presence of:of org
Mathematics      name these mathematical:product of two:name
Earth Science    of the paleozoic:d double prime:of this mine
Science          is broken down:this compound s:the photoelec
Astronomy        a black hole:in the sky:of the universe:brig
```

Accuracy: 0.756

Top Features: Unigrams and Bigrams

History	khan:emperor:successor:occurred:minister:byz
Literature	play:poet:playwright:stories:author who:lite
Social Science	bull:economist:holiday:anthropologist:psycho
Geography	african:was built:bridge:red:sea:city s:this
Other	band:player:season:film:starring:team:oscar:
Biology	cellular:membrane:protein:cell:syndrome:chro
Fine Arts	composition:movement:painting:piano:painted:
Physics	matter:voltage:quark:physical:wavelength:phy
Chemistry	chemical:reacts:spectroscopy:ion:gases:molar
Mathematics	euler:matrix:curve:prime:mathematical:method
Earth Science	zone:forms:period:hot:mantle:boundary:discon
Science	largest:this protein:this number:an electron
Astronomy	found:object:galaxies:mass:distance:atmosphe

Top Features: Unigrams and Bigrams

```
History          khan:emperor:successor:occurred:minister:byz
Literature       play:poet:playwright:stories:author who:lite
Social Science   bull:economist:holiday:anthropologist:psycho
Geography        african:was built:bridge:red:sea:city s:this
Other            band:player:season:film:starring:team:oscar:
Biology          cellular:membrane:protein:cell:syndrome:chro
Fine Arts        composition:movement:painting:piano:painted:
Physics          matter:voltage:quark:physical:wavelength:phy
Chemistry        chemical:reacts:spectroscopy:ion:gases:molar
Mathematics      euler:matrix:curve:prime:mathematical:method
Earth Science    zone:forms:period:hot:mantle:boundary:discon
Science          largest:this protein:this number:an electron
Astronomy        found:object:galaxies:mass:distance:atmosphe
```

Accuracy: 0.803

Top Features: Unigrams, Bigrams, Trigrams

```
History          emperor:organized:occurred:russia:legislatic
Literature       play:poet:writer:playwright:literature:poems
Social Science   hindu:anthropologist:holiday:demand:economis
Geography        this river:of this river:was built:red:this
Other            team:player:film:code:movie:season:yr:oscar:
Biology          humans:syndrome:protein:cellular:gene:membra
Fine Arts        10 pointsname this:pointsname this:pointsnam
Physics          wavelength:materials:decay:material:voltage:
Chemistry        hydrogen:molecules:this equation:spectroscop
Mathematics      named after:problem:space:data:algebraic:met
Earth Science    zone:hot:boundary:areas:period:material:regi
Science          radiation:machine:momentum:this group:atom:s
Astronomy        supernova:telescope:mass:astronomical:radiat
```

Top Features: Unigrams, Bigrams, Trigrams

```
History          emperor:organized:occurred:russia:legislatic
Literature       play:poet:writer:playwright:literature:poems
Social Science   hindu:anthropologist:holiday:demand:economis
Geography        this river:of this river:was built:red:this
Other            team:player:film:code:movie:season:yr:oscar:
Biology          humans:syndrome:protein:cellular:gene:membra
Fine Arts        10 pointsname this:pointsname this:pointsnam
Physics          wavelength:materials:decay:material:voltage:
Chemistry        hydrogen:molecules:this equation:spectroscop
Mathematics      named after:problem:space:data:algebraic:met
Earth Science    zone:hot:boundary:areas:period:material:regi
Science          radiation:machine:momentum:this group:atom:s
Astronomy        supernova:telescope:mass:astronomical:radiat
```

Accuracy: 0.809

Top Features: Character *n*-grams

```
History          khan: case: post: colo: new :orth : died:ses
Literature       epic:novel: " : play :play : play: poem :poe
Social Science   rivers: rivers: idea: son : myth:deity: deit
Geography        red : pass :ers. :nation : is : lies: lies :
Other            nger :(rl: : (rl:: (rl: : game: you : use :
Biology          rine : dna : cell:plant: grow:ase, :some :ge
Fine Arts        solo:piece: opera : st. : opera:opera : chor
Physics          law :zero :less : time : due : can :ying :ti
Chemistry        mole: chemis: chemist: sulf:sion :nium : che
Mathematics      proof: be :proof: any : has : are : four: se
Earth Science    laye: form:rock : rock : or : know: known:la
Science          is\n : law :ine. : atom: 10 : are : this: n
Astronomy        lion :cloud: earth: eart: radi: its : mass:
```

Top Features: Character *n*-grams

History	khan: case: post: colo: new :orth : died:ses
Literature	epic:novel: " : play :play : play: poem :poe
Social Science	rivers: rivers: idea: son : myth:deity: deit
Geography	red : pass :ers. :nation : is : lies: lies :
Other	nger :(rl: : (rl:: (rl: : game: you : use :
Biology	rine : dna : cell:plant: grow:ase, :some :ge
Fine Arts	solo:piece: opera : st. : opera:opera : chor
Physics	law :zero :less : time : due : can :ying :ti
Chemistry	mole: chemis: chemist: sulf:sion :nium : che
Mathematics	proof: be :proof: any : has : are : four: se
Earth Science	laye: form:rock : rock : or : know: known:la
Science	is\n : law :ine. : atom: 10 : are : this: n
Astronomy	lion :cloud: earth: eart: radi: its : mass:

Accuracy: 0.808

What else? (Don't limit to this data)

What else? (Don't limit to this data)

- Wikipedia categories

What else? (Don't limit to this data)

- Wikipedia categories
- How questions are written (quotas per packet)

What else? (Don't limit to this data)

- Wikipedia categories
- How questions are written (quotas per packet)
- Syntax (paying attention to words after "this")

What else? (Don't limit to this data)

- Wikipedia categories
- How questions are written (quotas per packet)
- Syntax (paying attention to words after "this")
- Authors of questions

HW3

- Detecting spoilers
- Kaggle competition (individual level)
- Given code already works (don't change output format)
- Can only change features