Today: Some Data Science Techniques

In the last lecture, we learned some basic computing methods:
- divide and conquer,
- sorting,
- searching,
- hashing, and
- dynamic programming.

Today, let’s talk about some basic techniques for data science.

Many Useful Data Science Techniques

You have already heard about some of these ideas in our class, such as
- neural networks,
- machine learning, and
- linear regression.

But in order to leverage data
- for learning and decision making,
- there are many other useful tools.

Logistic Regression is a Popular Choice

Let’s start by generalizing regression.
There are many other functions to which we can fit data.

One popular choice is the logistic function to the right, which allows us to map (for example)
- continuous numbers (horizontal)
- into binary choices (vertical) or
- probabilities (alternate interpretation of vertical).
Least Squares Frequently Used to Measure Error

As we mentioned earlier, fitting data to any kind of curve requires a metric that can be minimized. Often, these metrics measure the square distance between the fit value and a data point (called using least squares).

The benefit of such a choice is that finding the “best” is then easy using calculus: slope is proportional to distance.

Classification: Choosing the Right Category or Value

Given a set of data with many features, one can use probabilistic reasoning based on Bayes’ Theorem to “classify” missing features of a new data point. (In other words, to pick the most likely value for the missing features).

Let’s start with a simple example—the game of craps—then develop a Naïve Bayes Classifier for a more complex example.

Craps: Make Your Point to Win

Craps is a game played with 2 dice.

On a roll of 4, 5, 6, 8, 9, or 10, that value becomes the “point” and must be rolled again before rolling a 7. Doing so is called making the point.

Use Data to Estimate Likelihoods

How likely is it that you make your point if you roll an 8?

One can calculate that answer, but imagine that we instead collect a bunch of data.
Data on Making the Point in Craps

<table>
<thead>
<tr>
<th>Point</th>
<th>Made</th>
<th>Count</th>
<th>Point</th>
<th>Made</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Y</td>
<td>4,120</td>
<td>4</td>
<td>N</td>
<td>8,240</td>
</tr>
<tr>
<td>5</td>
<td>Y</td>
<td>6,658</td>
<td>5</td>
<td>N</td>
<td>9,781</td>
</tr>
<tr>
<td>6</td>
<td>Y</td>
<td>9,703</td>
<td>6</td>
<td>N</td>
<td>11,391</td>
</tr>
<tr>
<td>8</td>
<td>Y</td>
<td>9,652</td>
<td>8</td>
<td>N</td>
<td>11,485</td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>6,682</td>
<td>9</td>
<td>N</td>
<td>9,884</td>
</tr>
<tr>
<td>10</td>
<td>Y</td>
<td>4,147</td>
<td>10</td>
<td>N</td>
<td>8,257</td>
</tr>
</tbody>
</table>

Compute Probability Using Data

Let's use symbols for the events:

- \( M \equiv \text{making the point} \)
- \( E \equiv \text{point is 8} \)

To find the answer, we apply Bayes' Theorem:

\[
\text{prob}(M \text{ AND } E) = \text{prob}(M \mid E) \times \text{prob}(E)
\]

Rewriting,

\[
\text{prob}(M \mid E) = \frac{\text{prob}(M \text{ AND } E)}{\text{prob}(E)}
\]

We can compute the two numbers on the right from our data.

Maximum Likelihood Estimation to the Rescue!

But those are just data.

- **How do we find** \( \text{prob}(M \text{ AND } E) \)?
- **Maximum likelihood estimation**!

\[
\text{prob}(M \text{ AND } E) = \frac{9,652}{100,000} = 0.09652
\]

is the most likely to have generated 9,652 instances out of 100,000.

Similarly, we estimate

\[
\text{prob}(E) = \frac{21,137}{100,000} = 0.21137
\]
Data Give a Good Estimate to the Correct Answer

Plugging in,
\[ \text{prob}(M | E) = \frac{\text{prob}(M \text{ AND } E)}{\text{prob}(E)} \]
\[ = \frac{0.09652}{0.21137} = 0.4566 \]

If we instead
\* calculate analytically, assuming fair dice,
\* we obtain 0.4545,
\* so that's a pretty good estimate.

Let's consider a more complex example.

What Color is the Car?

Jan just bought a new, imported sports car.
Pat can't afford to buy such a car.
Jan has offered to let Pat drive the car if Pat can guess the car's color.

Can Pat predict the car’s color?

Use Data to Predict the Car’s Color!

Pat collects data on car sales nearby.

The data include
\* new and used vehicles,
\* domestic and imported vehicles,
\* sports cars and SUVs,
\* four colors, and,
\* counts for each combination.

Data on Car Sales

<table>
<thead>
<tr>
<th>source</th>
<th>new vehicles</th>
<th>used vehicles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>color</td>
<td>color</td>
</tr>
<tr>
<td></td>
<td>type</td>
<td>type</td>
</tr>
<tr>
<td>sports car</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dom.</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SUV</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>sports car</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>import</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>SUV</td>
<td>9</td>
<td>11</td>
</tr>
</tbody>
</table>
Start Again with Bayes’ Theorem

Let’s define some events:
\[ R \equiv \text{a car is red}, \quad B \equiv \text{blue}, \quad W \equiv \text{white}, \quad G \equiv \text{green} \]
\[ N \equiv \text{a car is new} \]
\[ I \equiv \text{a car is imported} \]
\[ F \equiv \text{a car is fast (a sports car!)} \]
Bayes’ Theorem gives us:
\[
\text{prob(color|N AND I AND F)} = \frac{\text{prob(color AND N AND I AND F)}}{\text{prob(N AND I AND F)}}
\]

Denominator Identical for All Colors

We can get a similar expression for each of the four colors.
\[
\text{prob(color|N AND I AND F)} = \frac{\text{prob(color AND N AND I AND F)}}{\text{prob(N AND I AND F)}}
\]
We only care about the relative probabilities for a classifier. The denominator is the same—just a number. So we can ignore it in comparisons.

Apply Bayes’ Theorem Again to Break Apart

Let’s use \( \approx \) to denote proportionality and focus on red (R) for now:
\[
\text{prob(R|N AND I AND F)} \approx \frac{\text{prob(R AND N AND I AND F)}}{\text{prob(N AND I AND F)}}
\]
We can apply Bayes’ Theorem again on the right:
\[
\approx \text{prob(N AND I AND F | R)} \times \text{prob (R)}
\]

Assume Independence to Obtain Simple Factors

\[
\text{prob(R|N AND I AND F)} \approx \frac{\text{prob(N AND I AND F | R)} \times \text{prob (R)}}{\text{prob(N AND I AND F)}}
\]
To simplify, we can make the naïve assumption that factors N, I, and F are independent: in context, that exactly the same fraction of domestic and imported red cars are new, for example. We assume...
\[
\text{prob(N AND I AND F | R)} \approx \text{prob(N | R)} \times \text{prob(I | R)} \times \text{prob(F | R)}.
\]
Factors Can Be Estimated from Data

\[
\text{prob}(R | N \text{ AND } I \text{ AND } F) \approx \text{prob}(N | R) \times \text{prob}(I | R) \times \text{prob}(F | R) \times \text{prob}(R)
\]

The four values can be estimated from the data in the table for each of the four colors, (not just for R(ed)):

\[
\text{prob}(N | R) = \frac{\text{# of new red cars}}{\text{# red cars}} \\
\text{prob}(I | R) = \frac{\text{# of imported red cars}}{\text{# red cars}} \\
\text{prob}(F | R) = \frac{\text{# of red sports cars}}{\text{# red cars}} \\
\text{prob}(R) = \frac{\text{# of red cars}}{\text{total # cars}}
\]

Which Color Has the Highest Probability?

Plugging in,

\[
\text{prob}(R | N \text{ AND } I \text{ AND } F) \approx \frac{14}{27} \times \frac{24}{27} \times \frac{4}{27} \times \frac{27}{200} = 0.00922
\]

Similarly,

\[
\text{prob}(B | N \text{ AND } I \text{ AND } F) \approx \frac{11}{27} \times \frac{18}{27} \times \frac{16}{27} \times \frac{27}{200} = 0.0217
\]

\[
\text{prob}(W | N \text{ AND } I \text{ AND } F) \approx \frac{17}{83} \times \frac{65}{83} \times \frac{12}{83} \times \frac{83}{200} = 0.00962
\]

\[
\text{prob}(G | N \text{ AND } I \text{ AND } F) \approx \frac{22}{63} \times \frac{38}{63} \times \frac{8}{63} \times \frac{63}{200} = 0.00843
\]

Data on Car Sales

<table>
<thead>
<tr>
<th>new vehicles</th>
<th>used vehicles</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>type</td>
</tr>
<tr>
<td>dom. sports car</td>
<td>0</td>
</tr>
<tr>
<td>dom. SUV</td>
<td>4</td>
</tr>
<tr>
<td>import sports car</td>
<td>4</td>
</tr>
<tr>
<td>import SUV</td>
<td>9</td>
</tr>
</tbody>
</table>

The largest! So Pat guesses that Jan's car is blue.

What Color is the Car?

Can we predict the color? Yes! Have fun, Pat!
Naïve Bayes Classifier Assumes Independent Factors

That approach is called a **Naïve Bayes Classifier**.

The “naïve” part is
◦ the **assumption of independence of factors**,  
◦ which is **often not the case** in reality.

In the car data, independence held for each color (not across colors).

Use of Aggregated Statistics Helps Improve Answers

It’s worth noting
◦ that the **Naïve Bayes** Classifier
◦ **implicitly aggregates statistics**:  
◦ how many imported blue cars rather than  
◦ how many new, imported, blue sports cars.

As a result, it **may work better for small datasets**, even if the independence assumption does not hold, because the MLE estimates on aggregated statistics are better.

Specific Subcategories May be Noisy

From the data, for example,
\[
\begin{align*}
\text{prob}(N \text{ AND I AND F | R}) \times \text{prob}(R) &= \frac{4}{27} \times \frac{27}{200} = 0.0200 \\
\text{prob}(N \text{ AND I AND F | B}) \times \text{prob}(B) &= \frac{3}{27} \times \frac{27}{200} = 0.0150 \\
\text{prob}(N \text{ AND I AND F | W}) \times \text{prob}(W) &= \frac{4}{83} \times \frac{83}{200} = 0.0200 \\
\text{prob}(N \text{ AND I AND F | G}) \times \text{prob}(G) &= \frac{2}{64} \times \frac{63}{200} = 0.0100
\end{align*}
\]

Blue is the best guess for the model used to generate the data. These are both incorrect due to statistical variation.

Another Approach: K Nearest Neighbors (KNN)

If we have numerical data for a problem,
◦ we may instead want to **find**  
◦ “**nearby**” data points, such  
◦ as the **K Nearest Neighbors (KNN)**  
◦ for some value of K.

Given a new data point, its neighbors can
◦ **vote to pick a class / category**,  
◦ or their values can be **averaged to assign a value or probability** to the new point.
Making Loan Decisions with KNN

Imagine, for example, information about
• age,
• annual income,
• loan amount, and
• whether the person defaulted (failed to pay).

Where is Pat in the Dataset?

Pat comes to bank officer Jan.
Pat,
• age 30,
• with an annual income of $155k,
• wants to take out a loan of $175k
• to buy a home.

Should Jan approve the loan?
Let’s take a look.

Find the Five Data Points “Closest” to Pat’s Request

Let’s look around Pat.
distance = sqrt [(age – 30)^2 + (salary – 155)^2 + (loan – 175)^2]

Here are the five nearest neighbors.

Remember: Age is Third Dimension in These Data

What happened here?
That point (two people, actually) is 3 years younger than Pat.

But ... are years and k$ the same?
Dimensions Can be Weighted as Desired

We may want to put more (or less) weight on years and dollars.

I already did so by using \( k \) instead of \( \$ \).

We could add a factor for the age term.

Example: Multiply the Age Contribution by 10

For example,

\[
\text{distance} = \sqrt{10 \times (\text{age} - 30)^2 + (\text{salary} - 155)^2 + (\text{loan} - 175)^2}
\]

Now we get a different point...

Classification: Plurality of Votes Wins

So how does bank officer Jan decide?
If the points “vote,” we have
* 2 default, and
* 3 no default.

So Pat gets the loan.

Estimation: Average Values from Neighbors

Alternatively, Jan may estimate that Pat has a 40% (2/5) chance of defaulting.
Jan can then decide whether or not that chance is too high.
To improve accuracy, Jan can use larger \( K \).
Higher K Values Produce Smoother Answers

Larger values of K suggest that Pat may be a reasonably safe bet.

We need enough data, of course, for so many neighbors to be relevant.

<table>
<thead>
<tr>
<th>K</th>
<th>% Defaulted</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>40.0</td>
</tr>
<tr>
<td>10</td>
<td>30.0</td>
</tr>
<tr>
<td>15</td>
<td>20.0</td>
</tr>
<tr>
<td>20</td>
<td>15.0</td>
</tr>
<tr>
<td>25</td>
<td>12.0</td>
</tr>
<tr>
<td>30</td>
<td>13.3</td>
</tr>
<tr>
<td>35</td>
<td>11.4</td>
</tr>
<tr>
<td>40</td>
<td>12.5</td>
</tr>
<tr>
<td>45</td>
<td>15.6</td>
</tr>
<tr>
<td>50</td>
<td>14.0</td>
</tr>
</tbody>
</table>

See the Shape of the Data?

Recorded dimensions may also benefit from other types of transformations.

For example, you may notice that default rates go up with loan / income.

Let’s normalize loan by income...

Distance Equation Not Always Easy to Choose

Now the points look more uniformly distributed.

We need to choose new weights, though.

Different Types of Points May be Denser or Sparser

Sometimes, one type of data may be much more common than others.

Here, for example, few of the data represent younger borrowers (blue or default—see circles).
Data Can be Replaced by Clustered Points

In such cases,
° data can be replaced with clustered points
° such that the density of each type is equal,
° preventing KNN from being swamped
° with effectively equivalent points of one type.

Curse of Dimensionality: Too Many in Real Data

Here, we chose a problem with only three dimensions.
Real data may have hundreds of dimensions.
The problem is called the curse of dimensionality.
Before applying an approach such as KNN, it’s useful to reduce the number of dimensions.

Principal Component Analysis (PCA) Reduces Dimensions

One technique for doing so is called Principal Component Analysis (PCA).
Conceptually, we find an ellipse that bounds the data.
The largest dimension of the ellipse becomes the first dimension for the data.
We then select orthogonal dimensions one at a time until we have the number we want.
In this way, we maximize variance of the data in the new dimensions.

Example of PCA with Few Dimensions (2 to 1)

For example, if we wanted to reduce these two dimensions into one, we could draw an ellipse and then take the dimension shown as the one dimension.
Points are projected onto this dimension.
PCA Based on Linear Algebra

PCA, of course, is not accomplished by hand drawings, but with linear algebra based on the full dataset.

Replace Human Decisions for KNN with ML

Today, many of the decisions that we discussed:
- how many dimensions to include,
- transformations to apply,
- weights to use (or even the functional form for measuring distance, and
- the value of K to select,
can be parametrized and trained using machine learning, which sidesteps the need for a human to make choices.

Other Techniques from “Top 10” Lists

A few other techniques that you may want to learn about in the future:
- Decision trees / CART (classification and regression trees)
- Support vector machines
- Ensemble methods
- Unsupervised learning (clustering); apriori, k-means
- Singular value decomposition
- Reinforcement learning

Terminology You Should Know from These Slides

- logistic function
- least squares
- classification
- Naïve Bayes Classifier
- K Nearest Neighbors (KNN)
- curse of dimensionality
- Principal Component Analysis (PCA)
Concepts You Should Know from These Slides

- reason for use of squared distance metrics
- role of MLE in extracting probabilities from data
- assumptions, pros, and cons of Naïve Bayes Classifiers
- how to use Naïve Bayes Classifiers
- how to use K Nearest Neighbors
- choices necessary for KNN
- use of ML to make choices (in KNN, for example)