Challenge 4: Autoencoders

Unsupervised Machine Learning

Answer the following questions to the best of your ability. Be sure to show *all* code in-line, in addition to full written responses to each of the questions. Write in complete sentences where appropriate. **A complete submission is a single rendered PDF submitted to Canvas**.

As with the 3rd challenge, you will be using the 2016 American National Election Pilot Study data (`anes_2016.csv`). This challenge explores the *same goal* as the 3rd, but on the basis patterns learned from an autoencoder. As a reminder, your goal is to explore whether we can pick up on differences in a "social question space" along a dimension of *identity*. Identity is either "American" or "racial," on the basis of responses to a respective question. A naive expectation is that people's identities should be present in their responses to social questions.

Note: if you get stuck, remember to always look to the package help files, as well as class notes. You are welcome to continue using the h2o engine as we have in class, or you may use a different approach if you feel more comfortable with another architecture (there are several out there, e.g., keras/tensorflow).

The Social Issue Questions

Use the same 14 social issue questions as in the 3rd challenge. Question wording and response categories can be found in the Questionnaire.

- `vaccine` - "Do you favor, oppose, or neither favor nor oppose requiring children to be vaccinated in order to attend public schools?" (7 point from favor a great deal (1) to oppose a great deal (7))
- `autism` - "How likely or unlikely is it that vaccines cause autism?" (6 point from Extremely likely (1) to Extremely unlikely (6))
- `birthright_b` - "Do you favor, oppose, or neither favor nor oppose children of unauthorized immigrants automatically getting citizenship if they are born in this country?" (7 point from Favor a great deal (1) to Oppose a great deal (7))
- `forceblack` - "How often do you think police officers use more force than is necessary under the circumstances when dealing with BLACK people?" (5 point from Never (1) to Very often (5))
- `forcewhite` - "How often do you think police officers use more force than is necessary under the circumstances when dealing with WHITE people?" (5 point from Never (1) to Very often (5))
- `stopblack` - "How often do to think police officers stop BLACK people on the street without a good reason?" (5 point from Never (1) to Very often (5))
- `stopwhite` - "How often do to think police officers stop WHITE people on the street without a good reason?" (5 point from Never (1) to Very often (5))
- `freetrade` - "Do you favor, oppose, or neither favor nor oppose the U.S. making free trade agreements with other countries?" (7 point from Favor a great deal (1) to Oppose a great deal (7))
- `aa3` - "Do you favor, oppose, or neither favor nor oppose allowing universities to increase the number of underrepresented minority students studying at their schools by considering race along with other factors when choosing students?" (7 point from Favor a great deal (1) to Oppose a great deal (7))

- `warmdo` - "Do you think the federal government should be doing more about rising temperatures, should be doing less, or is it currently doing the right amount? (7 point from Should be doing a great deal more (1) to Should be doing a great deal less (7))
- `finwell` - "Do you think people's ability to improve their financial well-being is now better, worse, or the same as it was 20 years ago?" (7 point from A great deal better (1) to A great deal worse (7))
- `childcare` - "Do you favor an increase, decrease, or no change in government spending to help working parents pay for CHILD CARE when they can't pay for it all themselves?" (7 point from Increase a great deal (1) to Decrease a great deal (7))
- `healthspend` - "Do you favor an increase, decrease, or no change in government spending to help people pay for HEALTH INSURANCE when they can't pay for it all themselves?" (7 point from Increase a great deal (1) to Decrease a great deal (7))
- `minwage` - "Should the minimum wage be raised, kept the same, lowered but not eliminated, or eliminated altogether?" (4 point from Raised [1], Kept the same [2], Lowered [3], Eliminated [4])

The Task

Explore whether responses to these social questions are grouped along "identity"-specific lines. To do so, you will be building shallow and deep autoencoders to explore differences (if any) between those with weak or strong American identities, and then those with weak or strong racial identities, as it relates to responses to social issue questions.

Of note: You will need to use your weak/strong dichotomous features for each identity created from last week's challenge (#3 on self-organizing maps).

The Questions

The American Identity

1. (10 points) Build a shallow autoencoder with a single hidden layer consisting of 2 nodes on the full question space, but *not* including the dichotomous **American** identity feature. Then, extract the two "deep" features from the hidden layer and store these.
2. (10 points) Plot the two deep features against each other, with color conditioned by weak or strong **American** identity. Discuss the output in a *few sentences*. For example, do we see separation in the projection (question) space along senses of American identity or not? Why or why not do you think?
3. (10 points) Build a *deep* autoencoder with 3 hidden layers consisting of 2 nodes in each on the full question space, but again *not* including the dichotomous **American** identity feature. Then, extract the two deep features from the *third* hidden layer and store these.
4. (10 points) Plot the two deep features from the *3rd hidden layer* against each other, with color conditioned by weak or strong **American** identity. Discuss the output in a *few sentences*. For example, does deepening the network help to recover different patterns and/or clearer separation in the question space along this identity? Why or why not do you think? What do we gain and what do we lose by deepening the network?

The Racial Identity

5. (10 points) Build a shallow autoencoder with a single hidden layer consisting of 2 nodes on the full question space, but *not* including the dichotomous **race** identity feature. Then, extract the two "deep" features from the hidden layer and store these.

6. (10 points) Plot the two deep features against each other, with color conditioned by weak or strong **racial** identity. Discuss the output in a *few sentences*. For example, do we see separation in the projection (question) space along senses of racial-identity or not? Why or why not do you think?

7. (10 points) Build a *deep* autoencoder with 3 hidden layers consisting of 2 nodes in each on the full question space, but again *not* including the dichotomous **racial** identity feature. Then, extract the two deep features from the *third* hidden layer and store these.

8. (10 points) Plot the two deep features from the *3rd hidden layer* against each other, with color conditioned by weak or strong **racial** identity. Discuss the output in a *few sentences*. For example, does deepening the network help to recover different patterns and/or clearer separation in the question space along this identity? Why or why not do you think? What do we gain and what do we lose by deepening the network?

Self-Organizing Maps vs. Autoencoders

9. (20 points) Compare the patterns across these two identities - American and racial - to the patterns found from last week's challenge using self-organizing maps. Are the patterns similar or different across these techniques (SOM vs. AE)? Why do you think? What might the benefit be of picking one of these neural network-based approaches to dimension reduction over the other? What do we gain with such a choice and what do we lose? And so on. *7-10 well-constructed sentences should suffice.*