Predicting Disease Using Support Vector Machines: A Focus on Demographic, Lifestyle, and Dietary Factors

Mark Daza

DATA 5322 – Homework 2

Technical Background

What is SVM?

Support Vector Machine (SVM) is a classification approach. Initially developed in the computer science community, it quickly became one of the best "out of the box" classifiers.

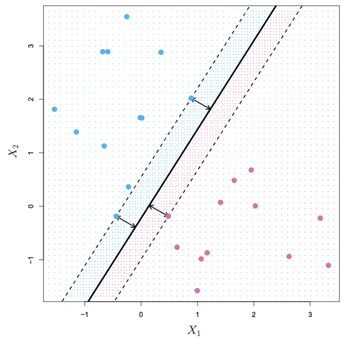

Figure 1. Maximal margin hyperplane [2]

We can think of SVM as a generalization of the concept of a ***maximal margin classifier*** (Figure 1). The ***maximal margin hyperplane*** is the solid line, while the ***margin*** is the distance from the solid line to either of the dashed lines. The points that lie on the dashed lines are the ***supporting vectors.*** A maximal margin classifier computes a line that separates the two classes, maximizing the distance between the line and the points.

SVM for Non-Linear Boundaries

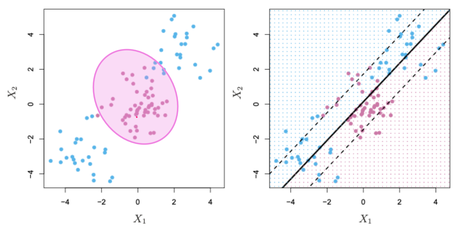

Figure 2. Example of non-linear boundaries [3]

While SVMs perform well with linear separation, real-world data often requires **non-linear boundaries** (Figure 2). By using a **kernel function,** SVMs can transform the input space and find complex decision boundaries.

Tuning Parameters in SVM

Support vector machines have important tuning parameters that affect how the decision boundary is created:

- **Cost (C)** controls the trade-off between maximizing the margin and minimizing misclassification.
- **Kernel:**
  - **Radial Kernel:** This allows curvier, even circular boundaries. Faraway points matter much less. The gamma tuning parameter adjusts the size of the sphere's influence.
  - **Polynomial Kernel**: Takes a linear relationship to the nth power and allows for curved boundaries —higher degree = more inflections.
  - **Linear Kernel**: Equivalent to a support vector classifier.

Dataset and Preprocessing

For this project, we used survey data from the **National Health Interview Survey (NHIS)[1]**. We focused on individuals who answered questions about health conditions, demographics, lifestyle, and eating habits.

Goal:

The goal was to predict whether a person has been diagnosed with **Cancer or Diabetes** based on selected features in the South region of the USA.

Figure 3. Southern United States

Data Cleaning and Processing:

Younger participants were excluded due to the low incidence of the diseases being studied. Rather than fully cleaning the dataset for all variables, we selectively cleaned variables relevant to our modeling to retain as much data as possible.

We grouped predictors into three main categories:

- **Demographics**: Sex, Education Level, Poverty Level, BMI, Height, Weight, Hours Worked
- **Physical Activity and Lifestyle**: Sleep Hours, Moderate and Vigorous Activity, Alcohol Use, Cigarette Use
- **Eating Habits**: Fruit Intake, Vegetable Intake, Juice Consumption, Soda, Fries, Pizza, and other dietary behaviors

Cleaning steps included removing implausible values and recoding unknown or missing categories where needed.

Modeling

We used **Support Vector Machines (SVMs)** to model the probability of disease diagnosis based on the selected predictors.

Separate models were trained for **demographic**, **physical activity/lifestyle**, and **eating habit** feature sets to assess which types of variables were most predictive.

Key modeling steps included:

- The target variables were a binary indicator for ever having a health condition diagnosis (1 = Yes, 0 = No).
- Scaling numeric predictors
- **Addressing class imbalance** by applying **class weights** during model training, giving greater importance to the disease class
- Splitting the data into training and testing sets, with a further **subset of 2000 observations** sampled from the training data for hyperparameter tuning to reduce computation time.
- Tuning hyperparameters (**cost** and **kernel type**) using cross-validation
- Exploring both **linear** and **non-linear kernels** (Radial and Polynomial)

Model performance was evaluated based on test errors and model accuracy.

Results

The following tables compare model performance based on cross-validation error, training and test error, accuracy, and misclassification rates across cancer and diabetes prediction categories.

Predictions for Cancer

Demographics:

Kernel	CV Error	Training Error	Test Error	Accuracy	Misclassification
Linear	53%	45.8%	45.6%	54.4%	45.6%
Radial	11.1%	0.5%	12.6%	87.4%	12.6%
Polynomial	43.1%	47.3%	51.6%	48.4%	51.6%

Physical Activity and Lifestyle:

Kernel	CV Error	Training Error	Test Error	Accuracy	Misclassification
Linear	41.9%	45.4%	46.9%	53.1%	46.9%
Radial	20.2%	18.3%	27.7%	72.2%	27.7%
Polynomial	68.6%	54.7%	55.6%	44.4%	55.6%

Eating Habits:

Kernel	CV Error	Training Error	Test Error	Accuracy	Misclassification
Linear	88%	88.2%	88.3%	11.7%	88.3%
Radial	11.8%	0.2%	12.1%	87.9%	12.1%
Polynomial	41.1%	43.5%	49.4%	39.1%	60.9%

Predictions for Diabetes

Demographics:

Kernel	CV Error	Training Error	Test Error	Accuracy	Misclassification
Linear	62.6%	64.6%	65.6%	34.5%	65.5%
Radial	12.5%	0.3%	11.6%	88.4%	11.6%
Polynomial	41.1%	43.5%	49.4%	50.6%	49.4%

Physical Activity and Lifestyle:

Kernel	CV Error	Training Error	Test Error	Accuracy	Misclassification
Linear	84.4%	87.3%	88.9%	11.1%	88.9%
Radial	20.4%	19.7%	28.2%	71.8%	28.2%
Polynomial	58%	64.3%	67.2%	32.8%	67.2%

Eating Habits:

Kernel	CV Error	Training Error	Test Error	Accuracy	Misclassification
Linear	88.1%	87.7%	87.9%	12.1%	87.9%
Radial	11.5%	0.3%	12.8%	87.2%	12.8%
Polynomial	49.4%	55.8%	62.6%	37.4%	62.6%

Discussion

During the analysis, several challenges were encountered that required various adjustments to the modeling approach, with SVMs:

Class Imbalance:

- The dataset had imbalanced classes for all the health conditions (cancer, heart disease, diabetes, heart attack, and stroke), which led to biased predictions in the initial models generated. SVM will try to correctly classify the class of the predictions, which can lead to a bias toward the majority class, giving us a misleadingly high overall accuracy. To address this, a class weight was applied, giving more importance to the minority class during model training (class weight – "1"=1, "2"=15).

Dataset Size:

- After cleaning the data, we had ~27,000 observations (only adults). Initial model training took an impractical amount of time to run,>30min per predictor group. This forced us to subset the data further. We decided to focus on the South region of the USA (region=3) due to its high prevalence of cancer (0.13) and diabetes (0.13).

Optimization Limitations:

- Conducting a full hyperparameter optimization for the different kernels was limited due to the high computational cost of cross-validation, especially for the most complex polynomial kernels. As a result, we limited our hyperparameter search to smaller sets of cost, gamma, and degree values. High cost (C) values above 100 were not evaluated.

Conclusion

Predictions for Cancer:

- Across the different predictor categories analyzed, the Radial kernel consistently outperformed both Linear and Polynomial Kernels.
- The Radial kernel achieved the highest accuracy rates in all categories, but there appears to be a stronger accuracy when using demographics (87.4%) and eating habits (87.9%) categories.

Predictions for Diabetes:

- Similar to the cancer models, Radial performed the best across all the prediction groups.
- Radial kernel showed the highest accuracy with demographic (88.4%) and eating habits (87.2%) categories.

Results for both cancer and diabetes predictions show that demographic and eating habit features were the strongest predictors for health disease in the South region of the US.

Citations:

[1] Lynn A. Blewett, Julia A. Rivera Drew, Miriam L. King, Kari C.W. Williams, Daniel Backman, Annie Chen, and Stephanie Richards. IPUMS Health Surveys: National Health Interview Survey, Version 7.4 [dataset]. Minneapolis, MN: IPUMS, 2024. https://doi.org/10.18128/D070.V7.4. Links to an external site.http://www.nhis.ipums.org

[2] Source: Adapted from James et al., 2023, Figure 9.3.

[3] Source: Adapted from James et al., 2023, Figure 9.8.