

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Generative Modelling

Generative modeling

- Logistic regression is an example of **descriptive modeling** of the data where we are interested $\Pr[y|X]$.
- A richer way (*at the cost of generative assumptions on the data*) is to model $\Pr[X|y]$ and $\Pr[y]$ (i.e., the joint distribution on predictors and response). This is known as **generative modeling** and is more informative. We use **Bayes rule** to find $\Pr[y|X] = \frac{\Pr[X|y] \cdot \Pr[y]}{\Pr[X]}$.
- Example: Suppose there is a test for a rare disease that only 1% of the 1 million population has. The test is 99% accurate (for both people with/without the disease). What is the probability that a random person who has tested positive for the disease indeed has the disease?

Generative modeling

- Logistic regression is an example of **descriptive modeling** of the data where we are interested $\Pr[y|X]$.

- A richer way (at the cost of generative assumptions on the data) is to model the joint distribution on predictors and response). This is known as **generative modeling** and is more

Likelihood and Prior

informative. We use **Bayes rule** to find Posterior $= \frac{\Pr[X|y] \cdot \Pr[y]}{\Pr[X]}$.

- Example: Suppose there is a test for a rare disease that only 1% of the 1 million population has. The test is 99% accurate (for both people with/without the disease). What is the probability that a random person who has tested positive for the disease indeed has the disease?

Generative modeling

- Logistic regression is an example of **descriptive modeling** of the data where we are interested $\Pr[y|X]$.
- A richer way (*at the cost of generative assumptions on the data*) is to model $\Pr[X|y]$ and $\Pr[y]$ (i.e., the joint distribution on predictors and response). This is known as **generative modeling** and is more informative. We use **Bayes rule** to find $\Pr[y|X] = \frac{\Pr[X|y] \cdot \Pr[y]}{\Pr[X]}$.
 - Example: Suppose there is a test for a rare disease that only 1% of the 1 million population has. The test is 99% accurate (for both people with/without the disease). What is the probability that a random person who has tested positive for the disease indeed has the disease?
- We can use this idea to build a classifier:
 - Use the data to estimate the priors and likelihoods, i.e., $\Pr[y]$, $\Pr[X|y]$
 - Use Bayes rule to calculate posterior and use this for prediction, i.e., $\Pr[y|X]$

Generative modeling

- Logistic regression is an example of **descriptive modeling** of the data where we are interested $\Pr[y|X]$.
- A richer way (*at the cost of generative assumptions on the data*) is to model $\Pr[X|y]$ and $\Pr[y]$ (i.e., the joint distribution on predictors and response). This is known as **generative modeling** and is more informative. We use **Bayes rule** to find
$$\Pr[y|X] = \frac{\Pr[X|y] \cdot \Pr[y]}{\Pr[X]}.$$
 - Example: Suppose there is a test for a rare disease that only 1% of the 1 million population has. The test is 99% accurate (for both people with/without the disease). What is the probability that a random person who has tested positive for the disease indeed has the disease?
- We can use this idea to build a classifier:
 - Use the data to estimate the priors and likelihoods, i.e., $\Pr[y]$, $\Pr[X|y]$
 - Use Bayes rule to calculate posterior and use this for prediction, i.e., $\Pr[y|X]$

Naïve Bayes Classifier:

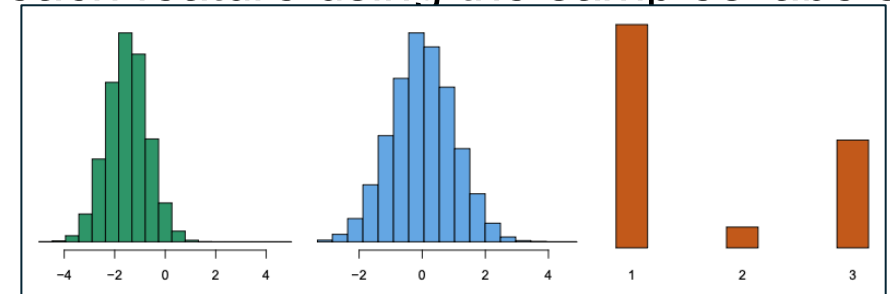
- $\Pr[y = k]$: Fraction of samples labeled k
- $\Pr[X|y = k] = \Pr[(X_1 = x_1, X_2 = x_2, X_3 = x_3)|y = k)]$
 $= \Pr[X_1 = x_1|y = k] \cdot \Pr[X_2 = x_2|y = k] \cdot \Pr[X_3 = x_3|y = k]$
 - Use the product distribution for multiple features. This assumes **independence** of features.
 - Compute a representative distribution for each feature using the samples labeled k .

Generative modeling

- Logistic regression is an example of **descriptive modeling** of the data where we are interested $\Pr[y|X]$.
- A richer way (*at the cost of generative assumptions on the data*) is to model $\Pr[X|y]$ and $\Pr[y]$ (i.e., the joint distribution on predictors and response). This is known as **generative modeling** and is more informative. We use **Bayes rule** to find
$$\Pr[y|X] = \frac{\Pr[X|y] \cdot \Pr[y]}{\Pr[X]}.$$
 - Example: Suppose there is a test for a rare disease that only 1% of the 1 million population has. The test is 99% accurate (for both people with/without the disease). What is the probability that a random person who has tested positive for the disease indeed has the disease?
- We can use this idea to build a classifier:
 - Use the data to estimate the priors and likelihoods, i.e., $\Pr[y]$, $\Pr[X|y]$
 - Use Bayes rule to calculate posterior and use this for prediction, i.e., $\Pr[y|X]$

Naïve Bayes Classifier:

- $\Pr[y = k]$: Fraction of samples labeled k
- $\Pr[X|y = k] = \Pr[(X_1 = x_1, X_2 = x_2, X_3 = x_3)|y = k)]$
$$= \Pr[X_1 = x_1|y = k] \cdot \Pr[X_2 = x_2|y = k] \cdot \Pr[X_3 = x_3|y = k]$$
 - Use the product distribution for multiple features. This assumes **independence** of features.
- Compute a representative distribution for each feature using the samples labeled k .



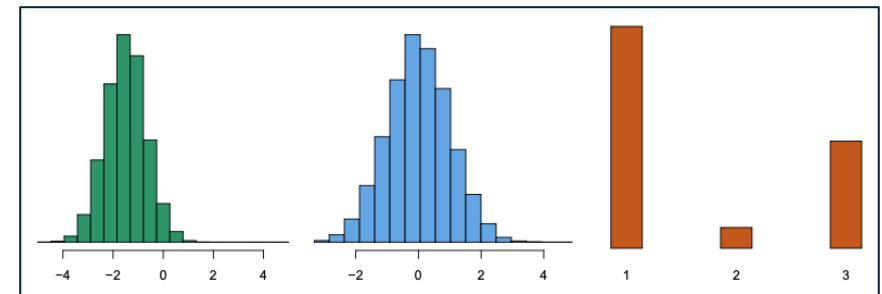
- Use **Kernel Density Estimator** (a smoothed histogram) for quantitative and density for qualitative variables.
- Use normal distribution with estimated parameters.

Generative modeling

- Logistic regression is an example of **descriptive modeling** of the data where we are interested $\Pr[y|X]$.
- A richer way (*at the cost of generative assumptions on the data*) is to model $\Pr[X|y]$ and $\Pr[y]$ (i.e., the joint distribution on predictors and response). This is known as **generative modeling** and is more informative. We use **Bayes rule** to find
$$\Pr[y|X] = \frac{\Pr[X|y] \cdot \Pr[y]}{\Pr[X]}.$$
 - Example: Suppose there is a test for a rare disease that only 1% of the 1 million population has. The test is 99% accurate (for both people with/without the disease). What is the probability that a random person who has tested positive for the disease indeed has the disease?
- We can use this idea to build a classifier:
 - Use the data to estimate the priors and likelihoods, i.e., $\Pr[y]$, $\Pr[X|y]$
 - Use Bayes rule to calculate posterior and use this for prediction, i.e., $\Pr[y|X]$

Naïve Bayes Classifier:

- $\Pr[y = k]$: Fraction of samples labeled k
- $\Pr[X|y = k] = \Pr[(X_1 = x_1, X_2 = x_2, X_3 = x_3)|y = k]$
 $= \Pr[X_1 = x_1|y = k] \cdot \Pr[X_2 = x_2|y = k] \cdot \Pr[X_3 = x_3|y = k]$
 - Use the product distribution for multiple features. This assumes **independence** of features.
 - Compute a representative distribution for each feature using the samples labeled k .



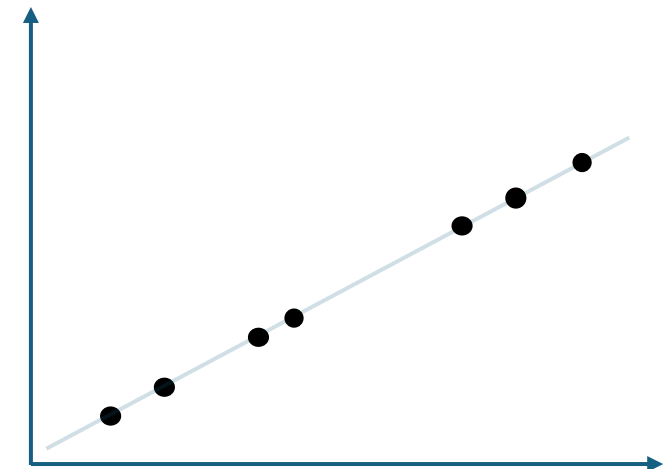
- Use **Kernel Density Estimator** (a smoothed histogram) for quantitative and density for qualitative variables.
- Use normal distribution with estimated parameters.

LDA, QDA: Drop independence

Unsupervised Learning

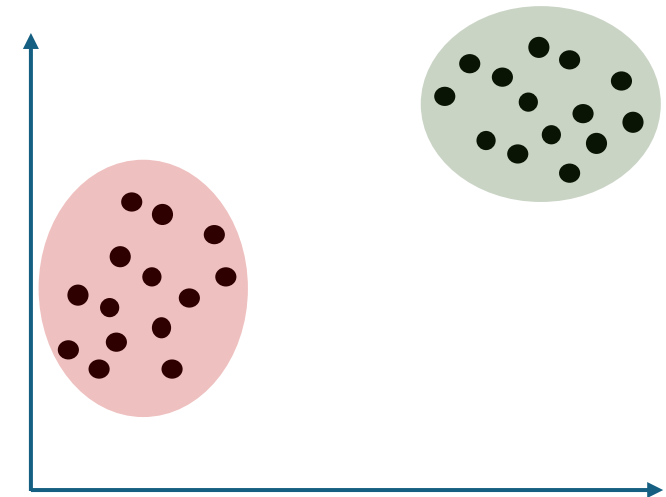
Classification

- Course content review:
 - Hypothesis testing: *Is the observation due to chance or otherwise?*
 - Parameter Estimation: *Can we estimate some parameter of the population from a sample?*
 - Supervised Learning: *Can we predict a response variable (regression if numerical and classification if categorical) using prediction variables?*
 - Unsupervised learning:
 - There is no response variable. We only have a set of features measured on n observations.
 - So, what are the data analysis questions of interest?
 - *Are there interesting patterns in the data?*
 - *Can we partition the observations into meaningful groups?*
 - *Can we compress the data in some meaningful way?*



Classification

- Course content review:
 - Hypothesis testing: *Is the observation due to chance or otherwise?*
 - Parameter Estimation: *Can we estimate some parameter of the population from a sample?*
 - Supervised Learning: *Can we predict a response variable (regression if numerical and classification if categorical) using prediction variables?*
 - Unsupervised learning:
 - There is no response variable. We only have a set of features measured on n observations.
 - So, what are the data analysis questions of interest?
 - *Are there interesting patterns in the data?*
 - *Can we partition the observations into meaningful groups?*
 - *Can we compress the data in some meaningful way?*



Classification

- Course content review:

- Hypothesis testing: *Is the observation due to chance or otherwise?*
- Parameter Estimation: *Can we estimate some parameter of the population from a sample?*
- Supervised Learning: *Can we predict a response variable (regression if numerical and classification if categorical) using prediction variables?*
- Unsupervised learning:
 - There is no response variable. We only have a set of features measured on n observations.
 - So, what are the data analysis questions of interest?
 - *Are there interesting patterns in the data?*
 - *Can we partition the observations into meaningful groups?*
 - *Can we compress the data in some meaningful way?*

- Suppose many people answer a personality questionnaire with items like:
"I enjoy being around people.", *"I keep my workspace organized."*, *"I worry a lot."*, *"I like trying new things."*, *"I am considerate of others."*
Each person gives numerical responses, say from 1 to 5.
- This produces a data matrix: $X \in \mathbb{R}^{n \times p}$ where n = number of people, and p = number of questionnaire items.
- Usually, many items are correlated. For example, several sociability-related items may move together. There may be a small number of linear combinations of the items that explain most of the variation.
- Suppose we have 4 questionnaire items: x_1 : *talkative*; x_2 : *enjoys parties*; x_3 : *organized*; x_4 : *follows schedules*
- And we can explain all observations as a linear combination of the variables Z_1 and Z_2 below.
$$Z_1 = 0.7 x_1 + 0.7 x_2 + 0.1 x_3 + 0.1 x_4$$
$$Z_2 = 0.1 x_1 + 0.1 x_2 + 0.7 x_3 + 0.7 x_4$$
- Then Z_1 looks like a sociability/extraversion dimension, and Z_2 looks like an orderliness/conscientiousness dimension.
- A person's responses can then be converted into numerical component scores.

End