

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

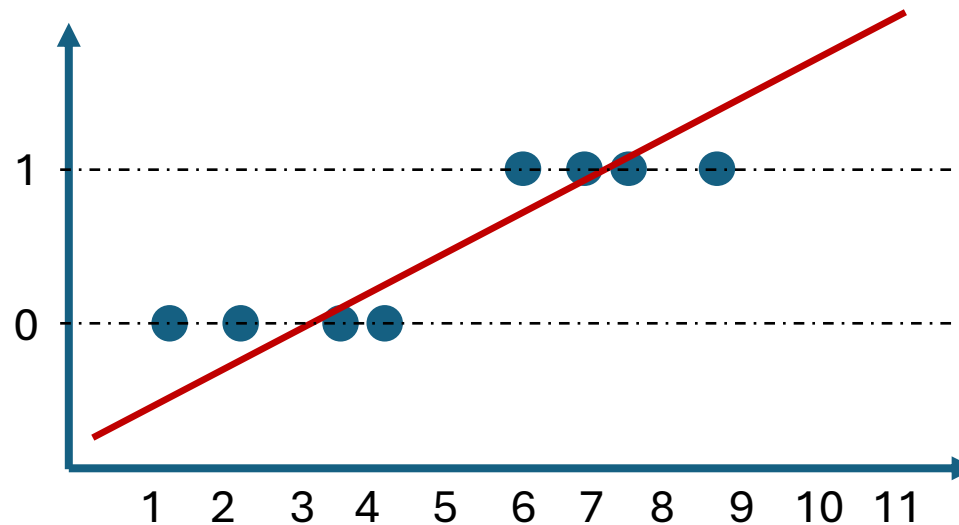
Logistic Regression

Logistic Regression

- Key question: *Can we use linear regression to predict a categorical response/target variable?*
 - *In principle, yes, but we need to adapt for better interpretation. Consider a 0/1 response variable along with a single numerical predictor. What does the regression line look like?*

Logistic Regression

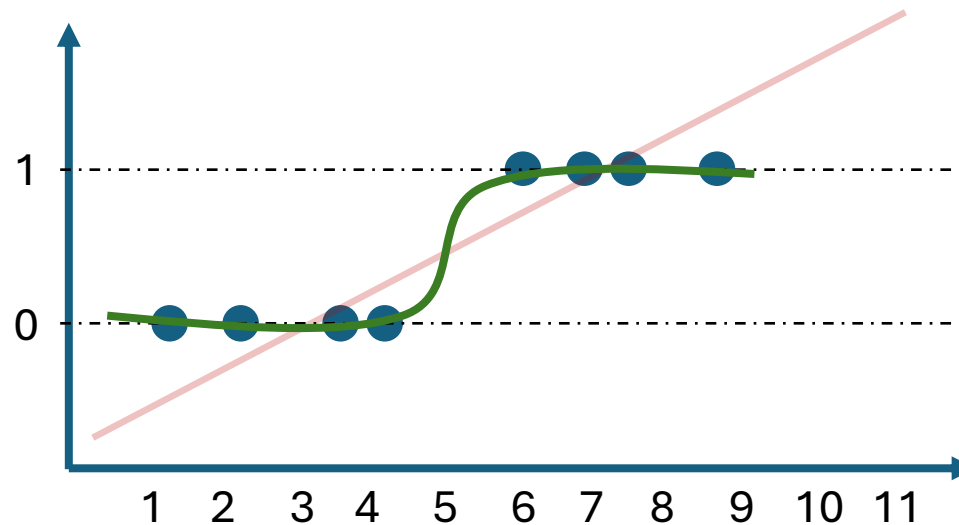
- Key question: *Can we use linear regression to predict a categorical response/target variable?*
 - *In principle, yes, but we need to adapt for better interpretation. Consider a 0/1 response variable along with a single numerical predictor. What does the regression line look like?*



- Issues:
 - *The regression line does not seem to be the best fitting curve through the data. Can you suggest a better curve?*

Logistic Regression

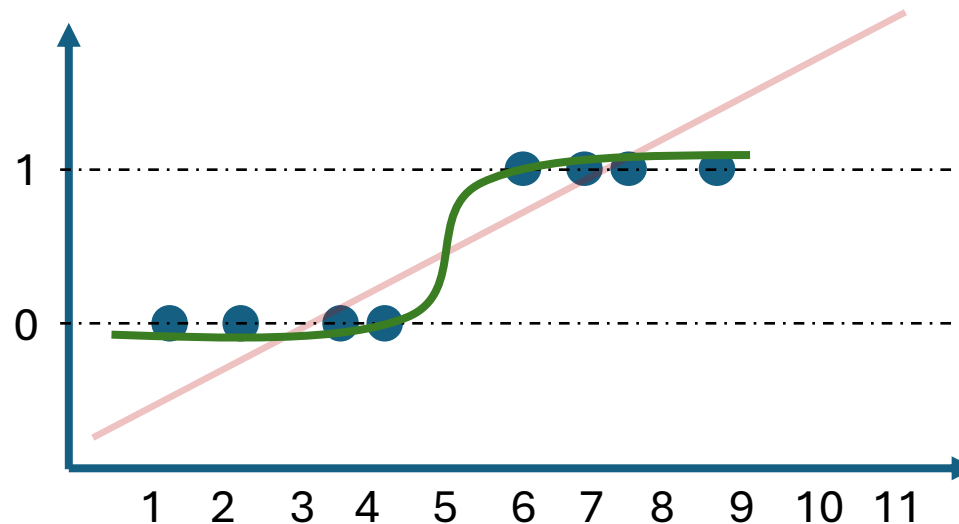
- Key question: *Can we use linear regression to predict a categorical response/target variable?*
 - *In principle, yes, but we need to adapt for better interpretation. Consider a 0/1 response variable along with a single numerical predictor. What does the regression line look like?*



- Issues:
 - *The regression line does not seem to be the best fitting curve through the data. Can you suggest a better curve?* $\sigma(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$

Logistic Regression

- Key question: *Can we use linear regression to predict a categorical response/target variable?*
 - *In principle, yes, but we need to adapt for better interpretation. Consider a 0/1 response variable along with a single numerical predictor. What does the regression line look like?*



- Issues:
 - *The regression line does not seem to be the best fitting curve through the data. Can you suggest a better curve?*
 $\sigma(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$ (the underlying function $\sigma(z) = \frac{1}{1 + e^{-z}}$ is called the **logistic function**)
 - *The output is supposed to be 0/1, but the linear predictor outputs arbitrary values. (Interpret value as probability)*

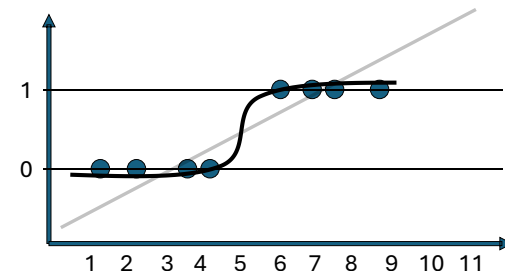
Logistic Regression

- Key question: *Can we use linear regression to predict a categorical response/target variable?*

- Yes. Fit the logistic curve $\sigma(\theta \cdot X)$.
- Question: What should be the **loss function** for evaluating the fit?
- Consider the following function:

$$L(\theta) = \prod_{i=1}^n p_i^{y_i} \cdot (1 - p_i)^{1-y_i}, \text{ where } p_i = \frac{1}{1+e^{-(\theta \cdot X_i)}}$$

- An interesting interpretation of $L(\theta)$: This is the **likelihood** (i.e., probability) of observing the data (i.e., (X, y)) assuming the logistic regression model!
- So, we find θ such that $L(\theta)$ gets maximized. Equivalently, we can maximize $l(\theta) = \log(L(\theta))$, which is called **log-likelihood** since it taking log converts products to sums which is easier to work with.



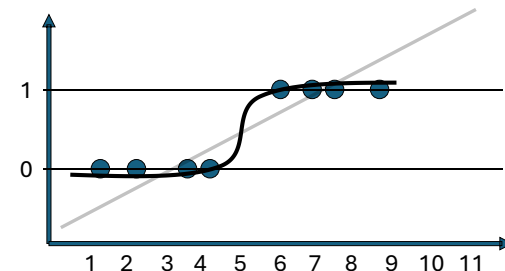
Logistic Regression

- Key question: *Can we use linear regression to predict a categorical response/target variable?*

- Yes. Fit the logistic curve $\sigma(\theta \cdot X)$.
- Question: What should be the **loss function** for evaluating the fit?
- Consider the following function:

$$L(\theta) = \prod_{i=1}^n p_i^{y_i} \cdot (1 - p_i)^{1-y_i}, \text{ where } p_i = \frac{1}{1+e^{-(\theta \cdot X_i)}}$$

- An interesting interpretation of $L(\theta)$: This is the **likelihood** (i.e., probability) of observing the data (i.e., (X, y)) assuming the logistic regression model!
- So, we find θ such that $L(\theta)$ gets maximized. Equivalently, we can maximize $\ell(\theta) = \log(L(\theta))$, which is called **log-likelihood** since it taking log converts products to sums which is easier to work with.
- Question: How do solve: Find θ such that $\ell(\theta)$ is maximized?



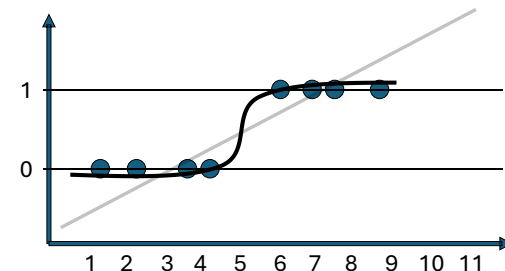
Logistic Regression

- Key question: Can we use linear regression to predict a categorical response/target variable?

- Yes. Fit the logistic curve $\sigma(\theta \cdot X)$.
- Question: What should be the **loss function** for evaluating the fit?
- Consider the following function:

$$L(\theta) = \prod_{i=1}^n p_i^{y_i} \cdot (1 - p_i)^{1-y_i}, \text{ where } p_i = \frac{1}{1+e^{-(\theta \cdot X_i)}}$$

- An interesting interpretation of $L(\theta)$: This is the **likelihood** (i.e., probability) of observing the data (i.e., (X, y)) assuming the logistic regression model!
- So, we find θ such that $L(\theta)$ gets maximized. Equivalently, we can maximize $\ell(\theta) = \log(L(\theta))$, which is called **log-likelihood** since it taking log converts products to sums which is easier to work with.
- Question: How do solve: Find θ such that $\ell(\theta)$ is maximized?
 - $\ell(\theta)$ is a nice concave function. So, we'll use Gradient ascent.



Logistic Regression

- Key question: Can we use linear regression to predict a categorical response/target variable?

- Yes. Fit the logistic curve $\sigma(\theta \cdot X)$.
- Question: What should be the **loss function** for evaluating the fit?

- Consider the following function:

$$L(\theta) = \prod_{i=1}^n p_i^{y_i} \cdot (1 - p_i)^{1-y_i}, \text{ where } p_i = \frac{1}{1+e^{-(\theta \cdot X_i)}}$$

- An interesting interpretation of $L(\theta)$: This is the **likelihood** (i.e., probability) of observing the data (i.e., (X, y)) assuming the logistic regression model!
- So, we find θ such that $L(\theta)$ gets maximized. Equivalently, we can maximize $\ell(\theta) = \log(L(\theta))$, which is called **log-likelihood** since it taking log converts products to sums which is easier to work with.
- Question: How do solve:
Find θ such that $\ell(\theta)$ is maximized?
 - $\ell(\theta)$ is a nice concave function. So, we'll use Gradient ascent.

- Question: What are the evaluation metrics for Yes/No prediction?

- **Confusion Matrix**:

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False Negative (FN)
Actual negative	False positive (FP)	True Negative (TN)

- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision: $\frac{TP}{TP+FP}$
- Recall: $\frac{TP}{TP+FN}$
- False Alarm Rate (FAR): $\frac{FP}{FP+TN}$
- Missed Detection Rate: $\frac{FN}{FN+TP}$
- F1 score: $2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Logistic Regression

- Key question: Can we use linear regression to predict a categorical response/target variable?

- Yes. Fit the logistic curve $\sigma(\theta \cdot X)$.
- Question: What should be the **loss function** for evaluating the fit?

- Consider the following function:

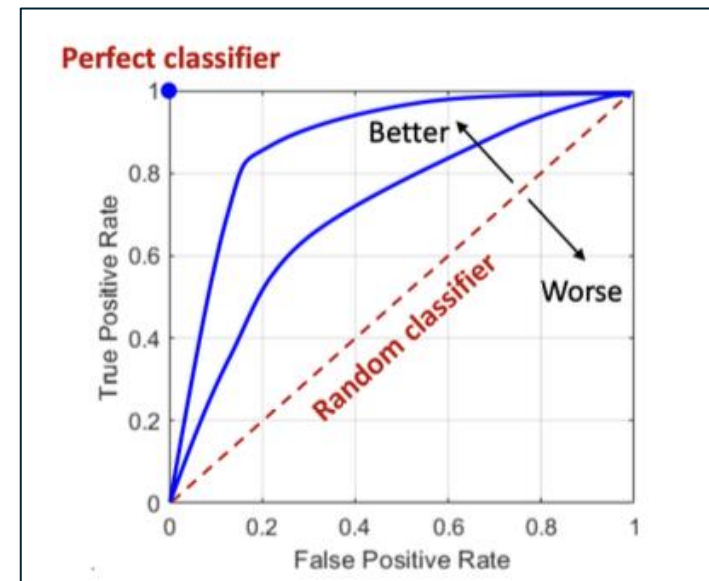
$$L(\theta) = \prod_{i=1}^n p_i^{y_i} \cdot (1 - p_i)^{1-y_i}, \text{ where } p_i = \frac{1}{1+e^{-(\theta \cdot X_i)}}$$

- An interesting interpretation of $L(\theta)$: This is the **likelihood** (i.e., probability) of observing the data (i.e., (X, y)) assuming the logistic regression model!
- So, we find θ such that $L(\theta)$ gets maximized. Equivalently, we can maximize $\ell(\theta) = \log(L(\theta))$, which is called **log-likelihood** since it taking log converts products to sums which is easier to work with.
- Question: How do solve:
Find θ such that $\ell(\theta)$ is maximized?
 - $\ell(\theta)$ is a nice concave function. So, we'll use Gradient ascent.

- Question: What are the evaluation metrics for Yes/No prediction?

- **Confusion Matrix**:

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False Negative (FN)
Actual negative	False positive (FP)	True Negative (TN)



ROC curve and AUC

Logistic Regression

- Key question: Can we use linear regression to predict a categorical response/target variable?

- Yes. Fit the logistic curve $\sigma(\theta \cdot X)$.
- Question: What should be the **loss function** for evaluating the fit?

- Consider the following function:

$$L(\theta) = \prod_{i=1}^n p_i^{y_i} \cdot (1 - p_i)^{1-y_i}, \text{ where}$$

$$p_i = \frac{1}{1 + e^{-(\theta \cdot X_i)}}$$

- An interesting interpretation of $L(\theta)$: This is the **likelihood** (i.e., probability) of observing the data (i.e., (X, y)) assuming the logistic regression model!
- So, we find θ such that $L(\theta)$ gets maximized. Equivalently, we can maximize $\ell(\theta) = \log(L(\theta))$, which is called **log-likelihood** since it taking log converts products to sums which is easier to work with.
- Question: How do solve:
Find θ such that $\ell(\theta)$ is maximized?
 - $\ell(\theta)$ is a nice concave function. So, we'll use Gradient ascent.

- Question: What are the evaluation metrics for Yes/No prediction?

- Use **Confusion Matrix**

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False Negative (FN)
Actual negative	False positive (FP)	True Negative (TN)

- Question: What do we do for non-binary categorical response variables?

- **Multinomial Logistic regression** using **softmax**

- When the number of categories $K > 2$

$$\Pr[y = k | X = x] = \frac{e^{-(\theta_k \cdot x)}}{\sum_{i=1}^K e^{-(\theta_i \cdot x)}}$$

- Rest of the details:

- Write maximization of the log-likelihood as an optimization problem and use Gradient ascent to solve the problem.

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 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:

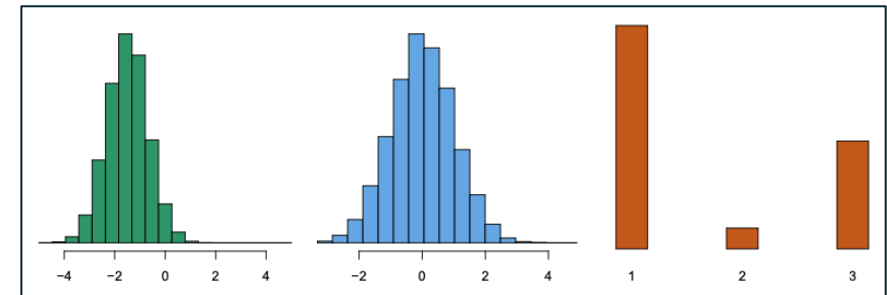
- $\Pr[y = k]$: Fraction of samples labeled k
- $\Pr[X|y = k] = \Pr[(X_1 = x_1, X_2 = x_2)|y = k]$
 $= \Pr[X_1 = x_1|y = k] \cdot \Pr[X_2 = x_2|y = k]$
 - Use the product distribution for multiple features. This assumes **independence** of features.
 - Compute a representative distribution for each feature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



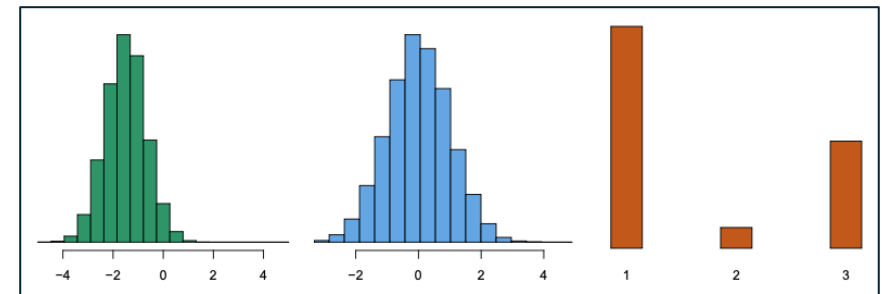
- Use **Kernel Density Estimator** (a smoothed histogram) for quantitative and density for qualitative variables.
- Use normal distribution with estimated parameters.

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
- $\Pr[X|y = k] = \Pr[(X_1 = x_1, X_2 = x_2)|y = k]$
 $= \Pr[X_1 = x_1|y = k] \cdot \Pr[X_2 = x_2|y = k]$
 - 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

End