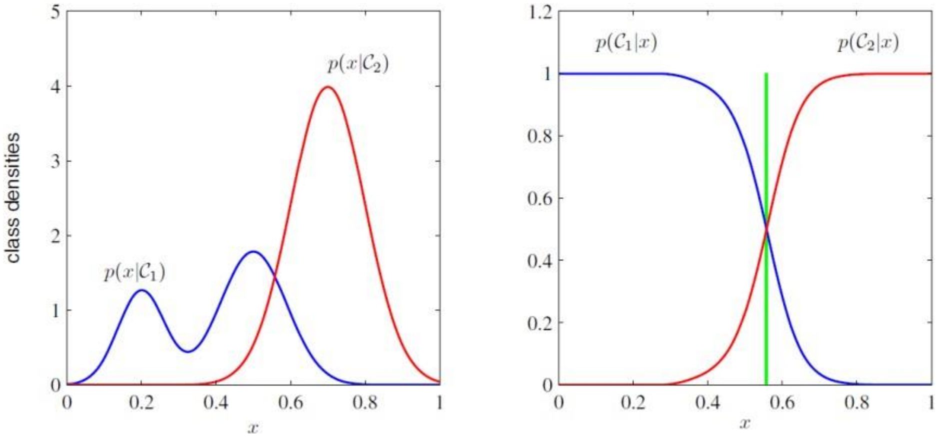
Naïve Bayes Classifer

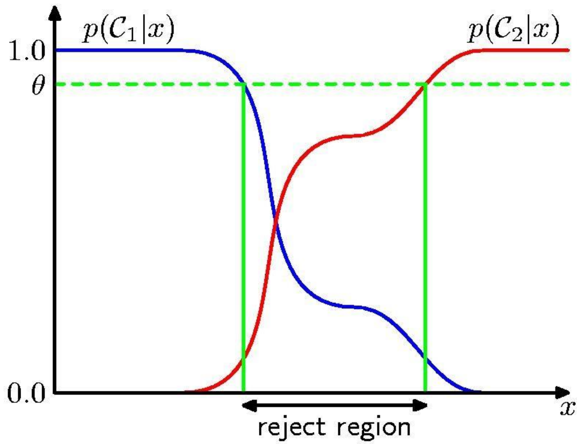
Inference and Decision

- Three approaches for classification
- **Generative model approach:**
  - Model $p(x, C_k) = p(C_k)p(x|C_k)$
  - Apply Bayes' Theorem for $p(C_k|x)$
  - Apply optimal decision criteria
- **Discriminative model approach:**
  - Model $p(C_k|x)$ directly
  - Apply optimal decision criteria
- **Discriminant function approach:**
  - Learn a function that maps each x to a class label directly from training data
  - No posterior probabilities!



Why separate Inference and Decision?

- Minimizing risk (loss matrix may change over time)
- Reject option
- Combining models (Popular Naïve Bayes classifier)
- And many more…



Problem Setting

$\chi \subseteq \Re^{p}$ is the input space

$X = \left(X_1, X_2, \cdots X_p \right)$ is a random variable describing the input

$\Upsilon \subseteq \Re$ or Γ is the output space

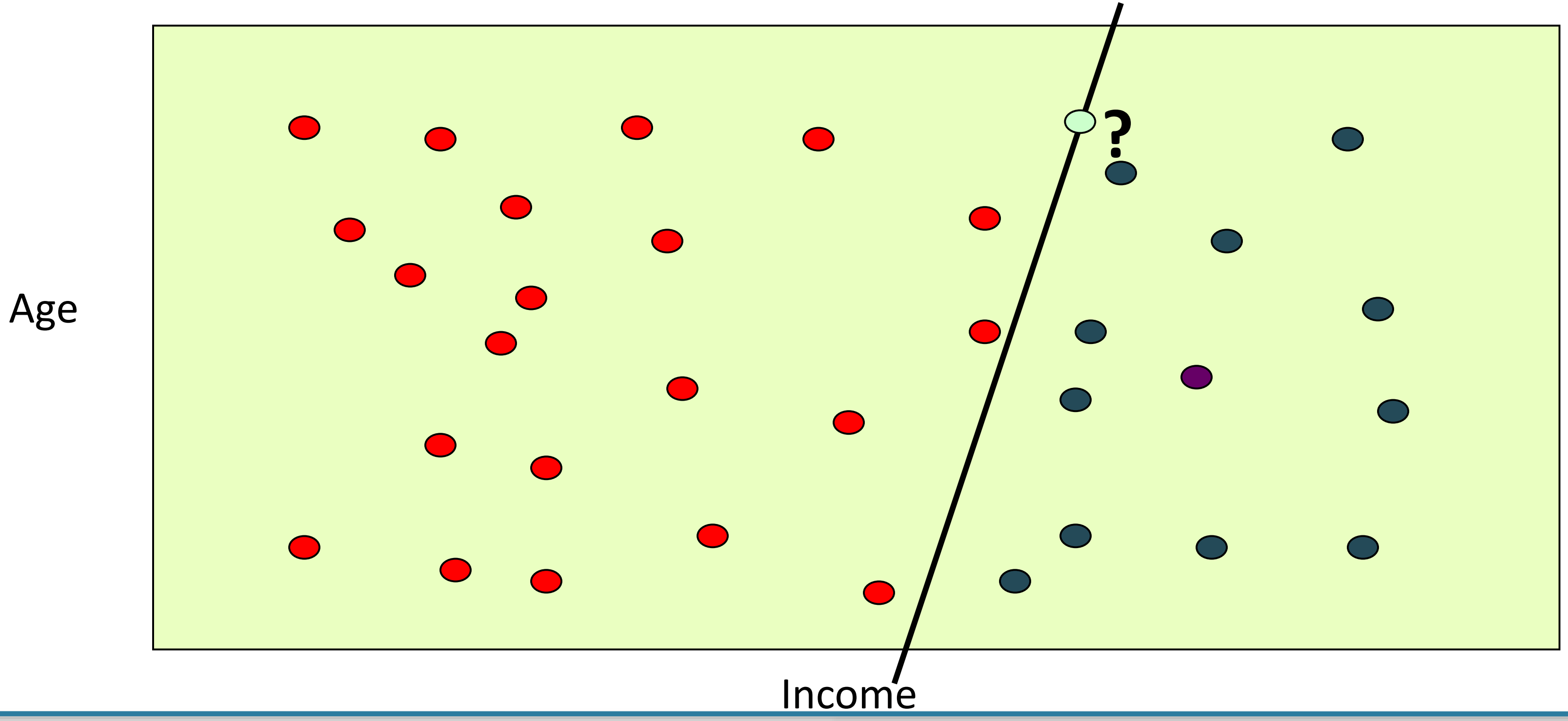
Y is a random variable describing the output

$p(X, Y)$ is the data distribution

$p(X, Y) = p(Y \middle| X) p(X)$

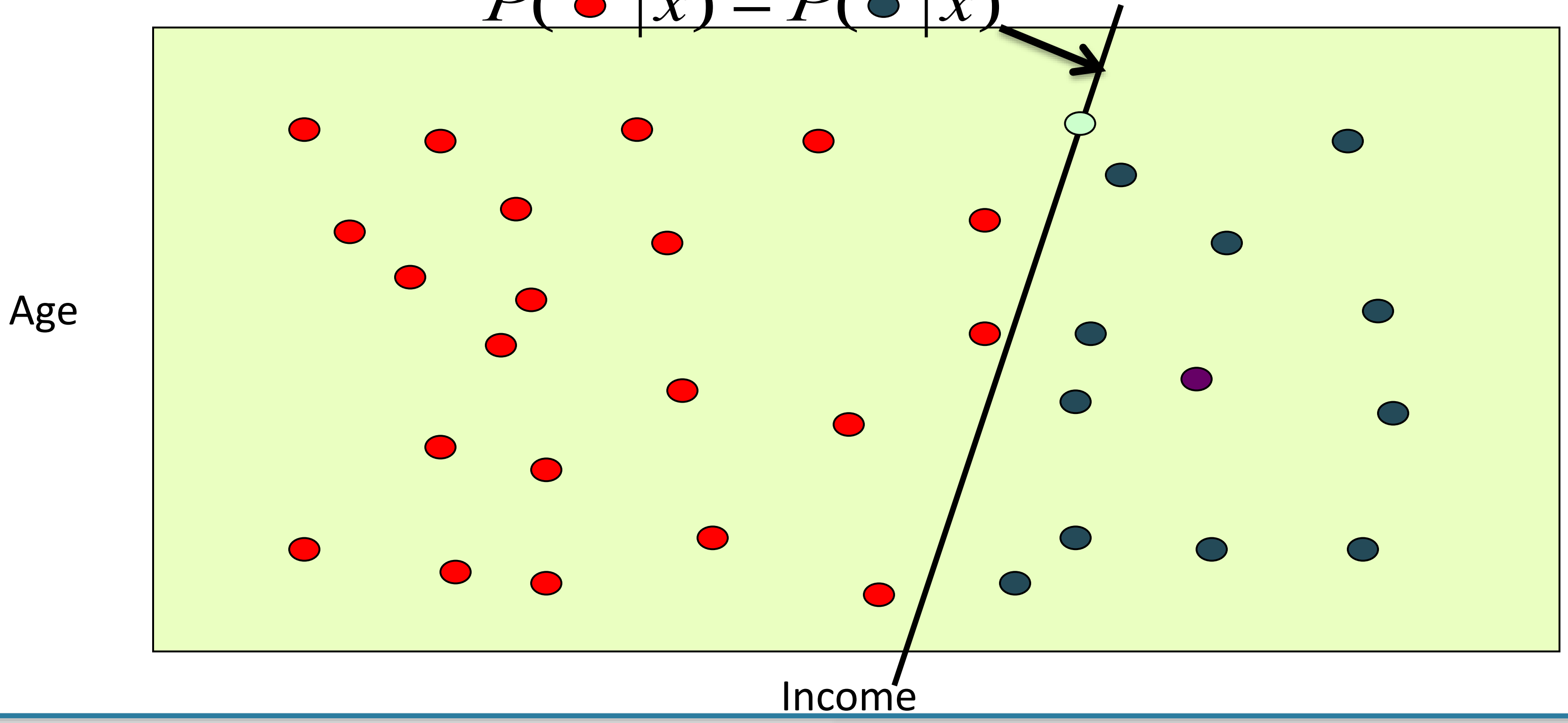
$p(Y \middle| x)$ is the predicted output probabilities given an input x

Interpretation



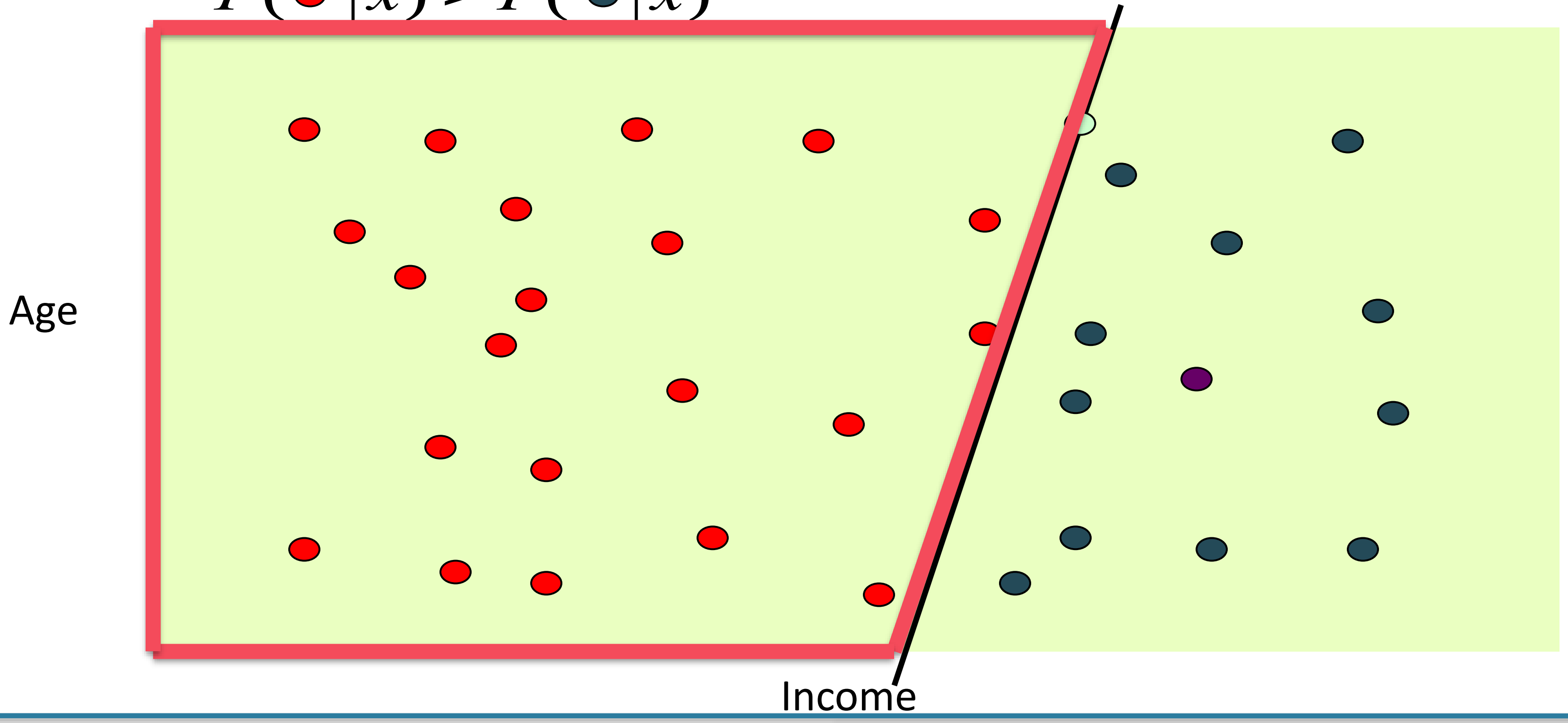
Separating Hyperplane

$$P(\bullet \mid x) = P(\bullet \mid x)$$



Separating Hyperplane

P(🔴 | *x*) > *P*(● | *x*)



Problem Setting

Assume that $Y = f(X) + e$,
where $E(e) = 0$, and independent of X.

Note that $f(x) = E\left(Y \middle| X = x\right)$
Goal is to find a f that minimizes expected prediction error (EPE).

For squared error loss, $\text{EPE}\left(f\right) = E\left(Y - f(X)\right)^2$
For classification - assume a $K\,´\;K$ loss matrix, $\mathbf{L}$.
The $\mathbf{L}_{ij}$ entry is the loss suffered by classifying class i as class j.
Typically use a $0 - 1$ loss function, where the off-diagonal entries of $\mathbf{L}$ are 1.

Problem Setting

Instead of assuming that $Y = f(X) + e$, one can directly model the $\mathrm{p}\left(G \middle| X\right)$ or $p(Y|X)$

This is sufficient to predict the output labels:

$\hat{G}\left(x\right)$=arg max$_{g}\, p\left(Y = g \middle| X = x\right)$

Depending on the assumptions we make on the form of p we get different classifiers.

Simplest of these is the Naive Bayes Assumption

Problem Setting

Instead of assuming that $Y = f(X) + e$, one can directly model the $p\left(G \middle| X\right)$ or $p(Y|X)$

This is sufficient to predict the output labels:

$$\hat{G}\left(x\right) = \text{arg max}_g\, p\left(Y = g \middle| X = x\right)$$

Depending on the assumptions we make on the form of p we get different classifiers.

Simplest of these is the Naive Bayes Assumption

Recall, Bayes theorem:

$$p\left(Y \middle| X\right) = \frac{p\left(X, Y\right)}{p\left(X\right)} = \frac{p\left(X \middle| Y\right) p\left(Y\right)}{p\left(X\right)}$$

Problem Setting

- Naïve Bayes Classifier
  - Assumption: The features are independent given the class labels

Problem Setting

- Naïve Bayes Classifier
  - Assumption: The features are independent given the class labels

$$\text{Independent: } p\left(X_1, X_2\right) = p\left(X_1\right) p\left(X_2\right)$$

$$\text{Conditionally independent: } p\left(X_1, X_2 \middle| Y\right) = p\left(X_1 \middle| Y\right) p\left(X_2 \middle| Y\right)$$

Naïve Bayes

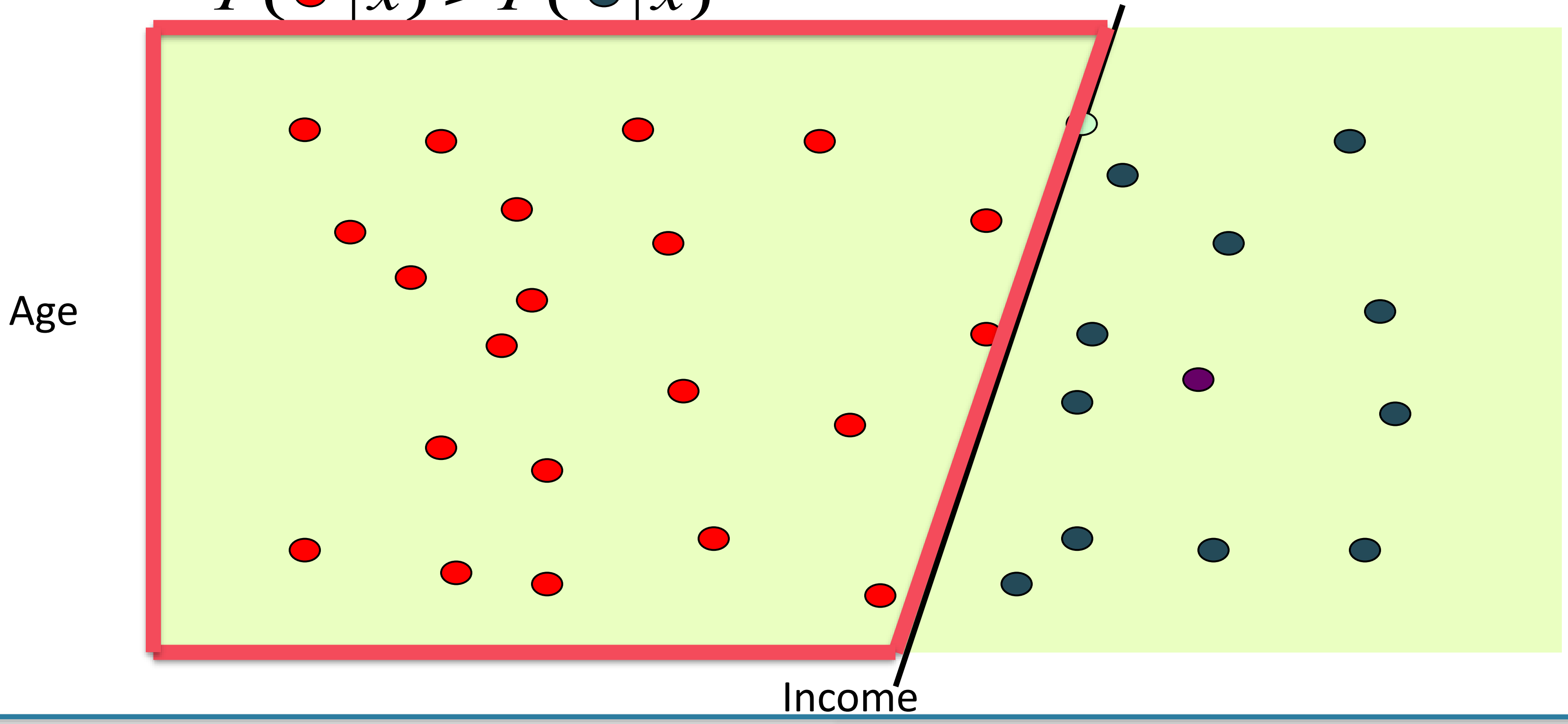
Naive Bayes assumption:

$$
\begin{aligned}
p(X|Y) &= p(X_1, X_2, \cdots, X_p|Y) \\
&= p(X_p|X_1, X_2, \cdots, X_{p-1}, Y)\, p(X_{p-1}|X_1, X_2, \cdots, X_{p-2}, Y) \cdots p(X_1|Y) \\
&= p(X_p|Y)\, p(X_{p-1}|Y) \cdots p(X_1|Y)
\end{aligned}
$$

$$
\begin{aligned}
p(Y|X) &= \frac{p(X_p|Y)\, p(X_{p-1}|Y) \cdots p(X_1|Y)\, p(Y)}{p(X)} \\
&\propto p(X_p|Y)\, p(X_{p-1}|Y) \cdots p(X_1|Y)\, p(Y)
\end{aligned}
$$

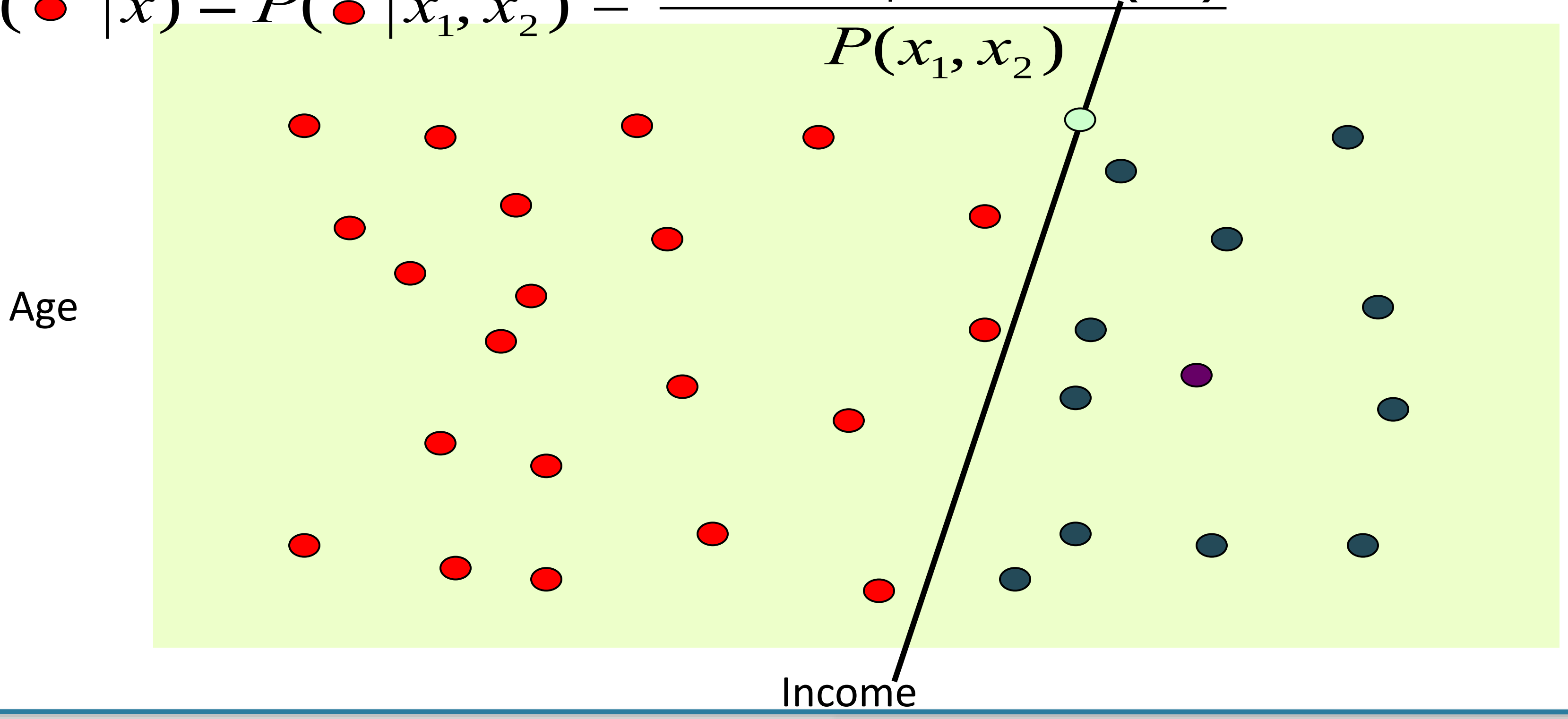
Separating Hyperplane

$$P(\text{🔴}\,|\,x) > P(\text{⚫}\,|\,x)$$



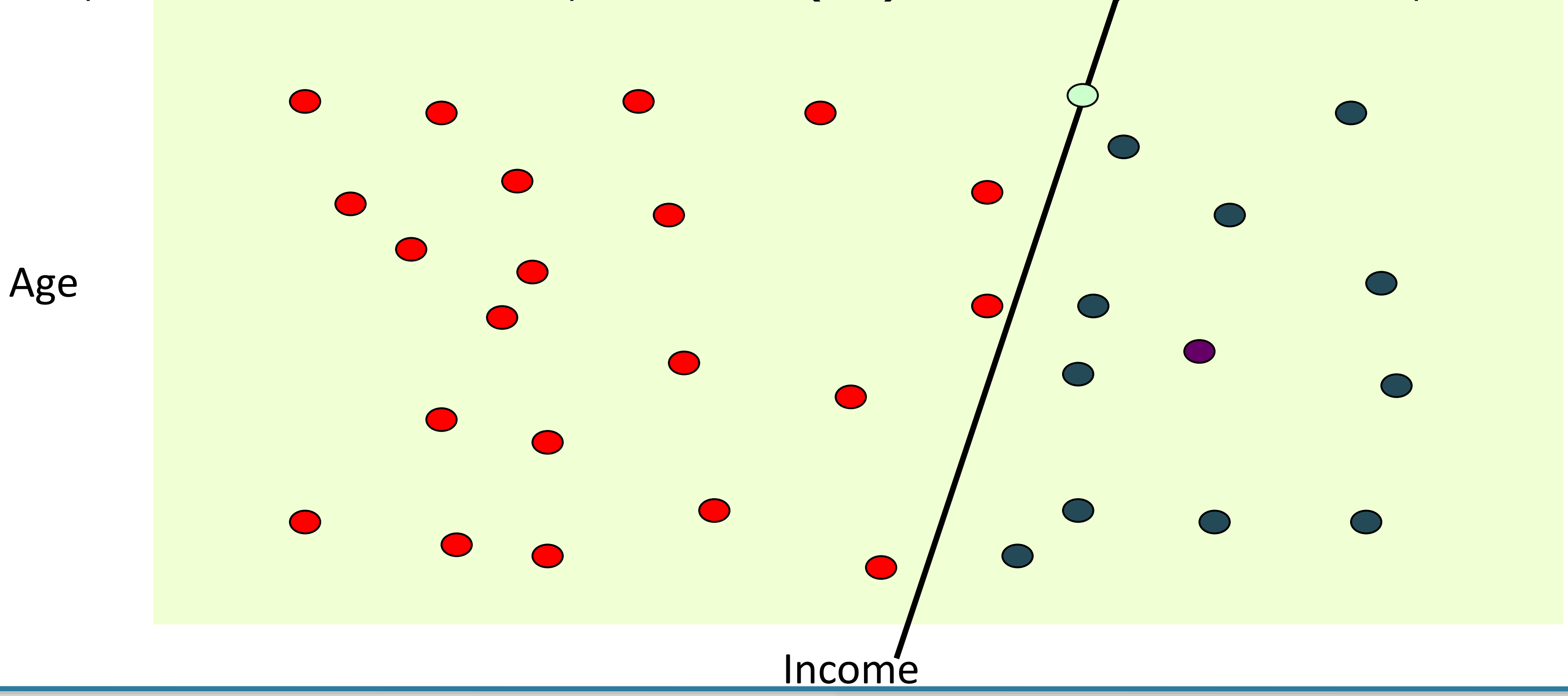
Bayes Theorem

$$P(\bullet\ |x) = P(\bullet\,|x_1, x_2) = \frac{P(x_1, x_2\,|\,\bullet\ )\acute{}\ \ P(\bullet)}{P(x_1, x_2)}$$



Naïve Bayes

$$P(\bullet \mid x_1, x_2)\ \mu\ P(x_1, x_2 \mid \bullet)\ \text{´}\ P(\bullet) = P(x_1 \mid \bullet)\ \text{´}\ P(x_2 \mid \bullet)\ \text{´}\ P(\bullet)$$

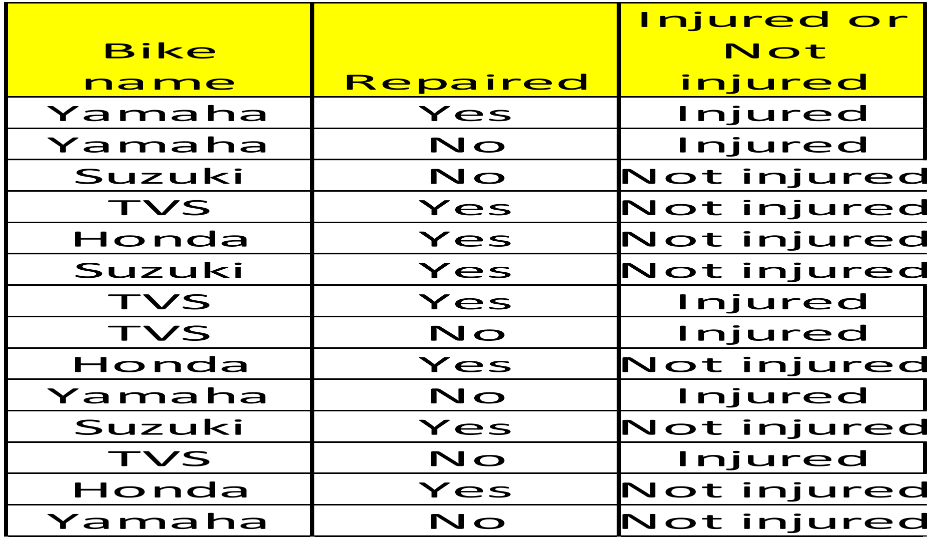


Naïve Bayes

- Assumption: The features are independent given the class labels
- Simple form for the probability distribution
- Not necessarily linear hyperplane ☺.
- Typically estimate by counting co-occurrences of feature value with class label
  - Maximum likelihood estimate
- Surprisingly powerful, especially in data with many features
  - High dimensional spaces

Understanding Bayes Theorem

Given the data of accident reports and status as injured or not injured of the person after the accident.

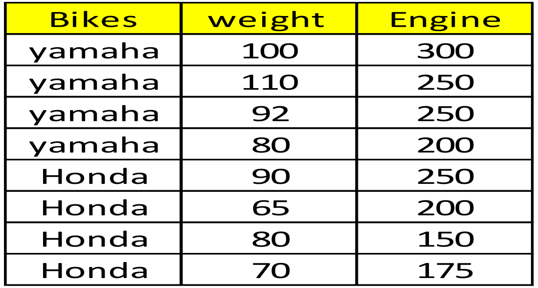


Bike name	Repaired	Injured or Not injured
Yamaha	Yes	Injured
Yamaha	No	Injured
Suzuki	No	Not injured
TVS	Yes	Not injured
Honda	Yes	Not injured
Suzuki	Yes	Not injured
TVS	Yes	Injured
TVS	No	Injured
Honda	Yes	Not injured
Yamaha	No	Injured
Suzuki	Yes	Not injured
TVS	No	Injured
Honda	Yes	Not injured
Yamaha	No	Not injured

Case: Yamaha and Not repaired

Classification through Bayes Theorem

Given data on bikes and their features



Bikes	weight	Engine
yamaha	100	300
yamaha	110	250
yamaha	92	250
yamaha	80	200
Honda	90	250
Honda	65	200
Honda	80	150
Honda	70	175

Predict the bike that was purchased from a given set of features,

Weight = 85 and engine = 250, Bike = ??

Where p(yamaha) = 0.5 and p(Honda) = 0.5

Assumptions

- Weight and engine are continuous variables
- Weight and engine are independent variables

Classification through Bayes Theorem

	Mean (weight)	Mean (Engine)	Variance (weights)	Variance (engine)
Yamaha(Y)	95.5	250	161	1666.66
Honda(H)	76.25	193.75	122.91	1822.91

Using Gaussian naïve Bayes,

P(Y/x(weight, engine)) = p(Y) *p(weight/Y)*p(engine/Y)*(1/p(x))

P(H/x) = p(H) *p(weight/H)*p(engine/H)*(1/p(x))

Using Gaussian distribution,

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

Probability	weight	engine
Yamaha	0.022331	0.009775
Honda	0.026361	0.003924

P(yamaha/x) > p(Honda/x) → YAMAHA

Continued…

When events model is discrete:

- Use frequency of every feature and class to estimate the likelihood and probabilities

When events model is continuous:

- Estimate the mean, variance for every feature of all training classes
- Use continuous models like Gaussian naive Bayes, Multinomial naive Bayes and Bernoulli naive Bayes