Learning classifiers using generative models

COMS 4721
Let \((\Omega, \mathbb{P})\) be a probability space. 
(\Omega \text{ is the sample space; } \mathbb{P} \text{ is the probability distribution.})
Let \((\Omega, \mathbb{P})\) be a probability space.
(\(\Omega\) is the sample space; \(\mathbb{P}\) is the probability distribution.)

For any events \(A, B \subseteq \Omega\),

\[
\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.
\]
Let \((\Omega, \mathbb{P})\) be a probability space. 
(\Omega\) is the sample space; \(\mathbb{P}\) is the probability distribution.)

For any events \(A, B \subseteq \Omega\),

\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)}.\]

**Bayes’ rule:**

\[
P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}.\]
Let $(\Omega, \mathbb{P})$ be a probability space.
($\Omega$ is the sample space; $\mathbb{P}$ is the probability distribution.)

For any events $A, B \subseteq \Omega$,

$$
\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.
$$

Bayes’ rule:

$$
\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B \mid A)}{\mathbb{P}(B)}.
$$

Let $E, H_0, H_1 \subseteq \Omega$. Conditioned on $E$, which of $H_0$ and $H_1$ is more probable?

Compare $\mathbb{P}(H_0) \cdot \mathbb{P}(E \mid H_0)$ to $\mathbb{P}(H_1) \cdot \mathbb{P}(E \mid H_1)$. 
Suppose result of medical test for disease is correct with probability 95%.
Also suppose disease is rare: any given person has disease with probability 1%.

**Question**: If test comes back positive for disease, is it more likely that you have disease or do not?
Suppose result of medical test for disease is correct with probability 95%.

Also suppose disease is rare: any given person has disease with probability 1%.

**Question:** If test comes back positive for disease, is it more likely that you have disease or do not?

\[
\begin{align*}
E &= \text{test comes back positive for disease} \\
H_0 &= \text{do not have disease} \\
H_1 &= \text{have disease}
\end{align*}
\]
Conditional probability example

Suppose result of medical test for disease is correct with probability 95%.
Also suppose disease is rare: any given person has disease with probability 1%.

**Question**: If test comes back positive for disease, is it more likely that you have disease or do not?

\[ E = \text{test comes back positive for disease} \]
\[ H_0 = \text{do not have disease} \]
\[ H_1 = \text{have disease} \]

\[ P(E \mid H_0) = 0.05 \]
\[ P(E \mid H_1) = 0.95 \]
Conditional probability example

Suppose result of medical test for disease is correct with probability 95%.
Also suppose disease is rare: any given person has disease with probability 1%.

**Question:** If test comes back positive for disease, is it more likely that you have disease or do not?

\[
E = \text{test comes back positive for disease} \\
H_0 = \text{do not have disease} \\
H_1 = \text{have disease} \\
P(E | H_0) = 0.05 \\
P(E | H_1) = 0.95 \\
P(H_0) = 0.99 \\
P(H_1) = 0.01
\]
Conditional probability example

Suppose result of medical test for disease is correct with probability 95%.
Also suppose disease is rare: any given person has disease with probability 1%.

Question: If test comes back positive for disease, is it more likely that you have disease or do not?

\[ E = \text{test comes back positive for disease} \]
\[ H_0 = \text{do not have disease} \]
\[ H_1 = \text{have disease} \]

\[ P(E \mid H_0) = 0.05 \]
\[ P(E \mid H_1) = 0.95 \]
\[ P(H_0) = 0.99 \]
\[ P(H_1) = 0.01 \]

Want to compare \( P(H_0 \mid E) \) to \( P(H_1 \mid E) \)
Suppose result of medical test for disease is correct with probability 95\%.

Also suppose disease is rare: any given person has disease with probability 1\%.

**Question**: If test comes back positive for disease, is it more likely that you have disease or do not?

\[
E = \text{ test comes back positive for disease}
\]
\[
H_0 = \text{ do not have disease}
\]
\[
H_1 = \text{ have disease}
\]
\[
P(E \mid H_0) = 0.05
\]
\[
P(E \mid H_1) = 0.95
\]
\[
P(H_0) = 0.99
\]
\[
P(H_1) = 0.01
\]

Want to compare \( P(H_0 \mid E) \) to \( P(H_1 \mid E) \), so compare

\[
P(H_0) \cdot P(E \mid H_0) = 0.99 \cdot 0.05 \quad \text{and} \quad P(H_1) \cdot P(E \mid H_1) = 0.01 \cdot 0.95.
\]
Let $X : \Omega \rightarrow \mathcal{X}$ be a random variable.
Let $X : \Omega \rightarrow \mathcal{X}$ be a random variable.

**Function of a random variable:**
For any function $g : \mathcal{X} \rightarrow \mathbb{R}$,

$$g(X) := g \circ X$$

is also a random variable:

$$g(X)(\omega) = g(X(\omega)).$$
Let $X : \Omega \to \mathcal{X}$ be a random variable.

**Function of a random variable:**
For any function $g : \mathcal{X} \to \mathbb{R}$,

$$g(X) := g \circ X$$

is also a random variable:

$$g(X)(\omega) = g(X(\omega)).$$

**Expected value:**

$$\mathbb{E}(g(X)) = \sum_{\omega \in \Omega} g(X(\omega)) \cdot \mathbb{P}(\omega)$$
Let \( X : \Omega \rightarrow \mathcal{X} \) be a random variable.

**Function of a random variable:**
For any function \( g : \mathcal{X} \rightarrow \mathbb{R} \),

\[
g(X) := g \circ X
\]

is also a random variable:

\[
g(X)(\omega) = g(X(\omega)).
\]

**Expected value:**

\[
\mathbb{E}(g(X)) = \sum_{\omega \in \Omega} g(X(\omega)) \cdot P(\omega)
\]

\[
= \sum_{\gamma} \gamma \cdot P(g(X) = \gamma).
\]
Review: conditional expectation

Let $X: \Omega \to \mathcal{X}$ and $Y: \Omega \to \mathbb{R}$ be random variables.
Let $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathbb{R}$ be random variables.

**Conditional expectation:**
For any $x \in \mathcal{X}$ such that $\mathbb{P}(X = x) > 0$:

$$\mathbb{E}[Y \mid X = x] := \sum_y y \cdot \mathbb{P}(Y = y \mid X = x).$$
Let $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathbb{R}$ be random variables.

**Conditional expectation:**
For any $x \in \mathcal{X}$ such that $\mathbb{P}(X = x) > 0$:

$$\mathbb{E}[Y \mid X = x] := \sum_y y \cdot \mathbb{P}(Y = y \mid X = x).$$

What is $\mathbb{E}[Y \mid X]$?
Let $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathbb{R}$ be random variables.

**Conditional expectation:**
For any $x \in \mathcal{X}$ such that $\mathbb{P}(X = x) > 0$:

$$
\mathbb{E}[Y \mid X = x] := \sum_y y \cdot \mathbb{P}(Y = y \mid X = x).
$$

What is $\mathbb{E}[Y \mid X]$? A random variable!

$$
\mathbb{E}[Y \mid X](\omega) = \mathbb{E}[Y \mid X = X(\omega)].
$$
Let $X: \Omega \rightarrow \mathcal{X}$ and $Y: \Omega \rightarrow \mathbb{R}$ be random variables.

**Conditional expectation:**
For any $x \in \mathcal{X}$ such that $\mathbb{P}(X = x) > 0$:

$$
\mathbb{E}[Y \mid X = x] := \sum_y y \cdot \mathbb{P}(Y = y \mid X = x).
$$

What is $\mathbb{E}[Y \mid X]$? A random variable!

$$
\mathbb{E}[Y \mid X](\omega) = \mathbb{E}[Y \mid X = X(\omega)].
$$

**Law of total expectation:**

$$
\mathbb{E}[\mathbb{E}[Y \mid X]] = \sum_{\omega \in \Omega} \mathbb{E}[Y \mid X = X(\omega)] \cdot \mathbb{P}(\omega)
$$

$$
= \sum_x \mathbb{E}[Y \mid X = x] \cdot \mathbb{P}(X = x) = \mathbb{E}(Y).
$$
Probability distribution $P$ over $\mathcal{X} \times \{0, 1\}$; let $(X, Y) \sim P$. 

The optimal classifier with smallest error rate (i.e., Bayes classifier) is $f^* (x) = \begin{cases} 0 & \text{if } \eta (x) \leq \frac{1}{2} \\ 1 & \text{if } \eta (x) > \frac{1}{2} \end{cases}$. 

Bayes classifier (for binary classification)
Bayes classifier (for binary classification)

- Probability distribution $P$ over $\mathcal{X} \times \{0, 1\}$; let $(X, Y) \sim P$.
- Think of $P$ as being comprised of two parts.
  1. Marginal distribution of $X$ (a distribution over $\mathcal{X}$).
  2. Conditional distribution of $Y$ given $X = x$, for each $x \in \mathcal{X}$:

$$
\eta(x) := P(Y = 1 \mid X = x).
$$
Bayes classifier (for binary classification)

Probability distribution $P$ over $\mathcal{X} \times \{0, 1\}$; let $(X, Y) \sim P$.

Think of $P$ as being comprised of two parts.

1. Marginal distribution of $X$ (a distribution over $\mathcal{X}$).
2. Conditional distribution of $Y$ given $X = x$, for each $x \in \mathcal{X}$:

$$\eta(x) := P(Y = 1 \mid X = x).$$

The optimal classifier with smallest error rate (i.e., Bayes classifier) is

$$f^*(x) = \begin{cases} 0 & \text{if } \eta(x) \leq 1/2 \\ 1 & \text{if } \eta(x) > 1/2 \end{cases}. $$
Bayes classifier (for binary classification)

Error rate of $f : \mathcal{X} \to \{0, 1\}$ can be written as 

$$\text{err}_P(f) = P(f(X) \neq Y) = \mathbb{E}[1\{f(X) \neq Y\}].$$
Bayes classifier (for binary classification)

Error rate of $f: \mathcal{X} \rightarrow \{0, 1\}$ can be written as

$$\text{err}_P(f) = P(f(X) \neq Y) = \mathbb{E}[\mathbf{1}\{f(X) \neq Y\}] .$$

Define $g: \mathcal{X} \rightarrow \mathbb{R}$ by

$$g(x) := \mathbb{E}[\mathbf{1}\{f(X) \neq Y\} | X = x] .$$
Bayes classifier (for binary classification)

Error rate of $f: \mathcal{X} \rightarrow \{0, 1\}$ can be written as

$$\text{err}_P(f) = P(f(X) \neq Y) = \mathbb{E}[1\{f(X) \neq Y\}].$$

Define $g: \mathcal{X} \rightarrow \mathbb{R}$ by

$$g(x) := \mathbb{E}[1\{f(X) \neq Y\} | X = x].$$

Then

$$g(x) = P(Y = 0 | X = x) \cdot 1\{f(x) \neq 0\} + P(Y = 1 | X = x) \cdot 1\{f(x) \neq 1\}.$$
Bayes classifier (for binary classification)

Error rate of $f : \mathcal{X} \to \{0, 1\}$ can be written as

$$\text{err}_P(f) = P(f(X) \neq Y) = \mathbb{E}[1\{f(X) \neq Y\}] .$$

Define $g : \mathcal{X} \to \mathbb{R}$ by

$$g(x) := \mathbb{E}[1\{f(X) \neq Y\} | X = x] .$$

Then

$$g(x) = P(Y = 0 | X = x) \cdot 1\{f(x) \neq 0\} + P(Y = 1 | X = x) \cdot 1\{f(x) \neq 1\}$$

$$= (1 - \eta(x)) \cdot 1\{f(x) = 1\} + \eta(x) \cdot 1\{f(x) = 0\} .$$
Bayes classifier (for binary classification)

Error rate of $f: \mathcal{X} \rightarrow \{0, 1\}$ can be written as

$$\text{err}_P(f) = P(f(X) \neq Y) = \mathbb{E}[\mathbbm{1}\{f(X) \neq Y\}] .$$

Define $g: \mathcal{X} \rightarrow \mathbb{R}$ by

$$g(x) := \mathbb{E}\left[\mathbbm{1}\{f(X) \neq Y\} \mid X = x\right] .$$

Then

$$g(x) = P(Y = 0 \mid X = x) \cdot \mathbbm{1}\{f(x) \neq 0\} + P(Y = 1 \mid X = x) \cdot \mathbbm{1}\{f(x) \neq 1\}$$

$$= (1 - \eta(x)) \cdot \mathbbm{1}\{f(x) = 1\} + \eta(x) \cdot \mathbbm{1}\{f(x) = 0\} .$$

What should $f(x)$ be so that $g(x)$ is as small as possible?
Error rate of $f: \mathcal{X} \to \{0, 1\}$ can be written as

$$\text{err}_P(f) = P(f(X) \neq Y) = \mathbb{E}[1\{f(X) \neq Y\}] .$$

Define $g: \mathcal{X} \to \mathbb{R}$ by

$$g(x) := \mathbb{E}[1\{f(X) \neq Y\} \mid X = x] .$$

Then

$$g(x) = P(Y = 0 \mid X = x) \cdot 1\{f(x) \neq 0\} + P(Y = 1 \mid X = x) \cdot 1\{f(x) \neq 1\}$$

$$= (1 - \eta(x)) \cdot 1\{f(x) = 1\} + \eta(x) \cdot 1\{f(x) = 0\} .$$

What should $f(x)$ be so that $g(x)$ is as small as possible?

Since $f(x) \in \{0, 1\}$, best to have

$$f(x) := \begin{cases} 0 & \text{if } \eta(x) \leq 1 - \eta(x) \\ 1 & \text{if } \eta(x) > 1 - \eta(x) \end{cases} .$$
Bayes classifier (for binary classification)

Error rate of $f: \mathcal{X} \rightarrow \{0, 1\}$ can be written as

$$\text{err}_P(f) = P(f(X) \neq Y) = \mathbb{E}[\mathbf{1}\{f(X) \neq Y\}] .$$

Define $g: \mathcal{X} \rightarrow \mathbb{R}$ by

$$g(x) := \mathbb{E}[\mathbf{1}\{f(X) \neq Y\} \mid X = x] .$$

Then

$$g(x) = P(Y = 0 \mid X = x) \cdot \mathbf{1}\{f(x) \neq 0\} + P(Y = 1 \mid X = x) \cdot \mathbf{1}\{f(x) \neq 1\}$$

$$= (1 - \eta(x)) \cdot \mathbf{1}\{f(x) = 1\} + \eta(x) \cdot \mathbf{1}\{f(x) = 0\} .$$

What should $f(x)$ be so that $g(x)$ is as small as possible?

Since $f(x) \in \{0, 1\}$, best to have

$$f(x) := \begin{cases} 
0 & \text{if } \eta(x) \leq 1 - \eta(x) \\
1 & \text{if } \eta(x) > 1 - \eta(x) 
\end{cases}$$

...which is the same as $f^*(x)$. □
Bayes classifier (for $K$-class classification)

- Probability distribution $P$ over $\mathcal{X} \times \{1, 2, \ldots, K\}$; let $(X, Y) \sim P$.
- Think of $P$ as being comprised of two parts.
  1. Marginal distribution of $X$ (a distribution over $\mathcal{X}$).
  2. Conditional distribution of $Y$ given $X = x$, for each $x \in \mathcal{X}$.
- Bayes classifier:

$$f^*(x) = \arg \max_{y \in \{1, 2, \ldots, K\}} P(Y = y \mid X = x).$$
By Bayes’ rule, Bayes classifier can be written as

\[
f^*(x) = \arg \max_{y \in \{1, 2, \ldots, K\}} P(Y = y) \cdot P(X = x \mid Y = y).
\]
By Bayes’ rule, Bayes classifier can be written as

\[ f^*(x) = \arg \max_{y \in \{1,2,\ldots,K\}} P(Y = y) \cdot P(X = x \mid Y = y). \]

Motivates thinking of \( P \) as being comprised of:

1. **Class priors** \( \pi_1, \pi_2, \ldots, \pi_K \in [0, 1] \), where \( \pi_y = P(Y = y) \).
2. **Class conditional distributions** \( P_1, P_2, \ldots, P_K \), where \( P_y \) is conditional distribution of \( X \) given \( Y = y \).
Structure of Bayes classifier

- By Bayes’ rule, Bayes classifier can be written as

\[ f^*(x) = \arg \max_{y \in \{1, 2, \ldots, K\}} P(Y = y) \cdot P(X = x \mid Y = y). \]

- Motivates thinking of \( P \) as being comprised of:

1. **Class priors** \( \pi_1, \pi_2, \ldots, \pi_K \in [0, 1] \), where \( \pi_y = P(Y = y) \).
2. **Class conditional distributions** \( P_1, P_2, \ldots, P_K \), where \( P_y \) is conditional distribution of \( X \) given \( Y = y \).

\[ \pi_1 = \frac{1}{5} \]
\[ \pi_2 = \frac{2}{5} \]
\[ \pi_3 = \frac{2}{5} \]

\( N(-2, 1/4) \)
\( N(0, 1) \)
\( N(1, 1/4) \)
In context of classification problems, a *generative model* is a statistical model $\mathcal{P}$ on $\mathcal{X} \times \{1, 2, \ldots, K\}$, where each $P_\theta \in \mathcal{P}$ is

$$P_\theta(x, y) = \pi_y \cdot P_{t_y}(x),$$

where the parameters are

$$\theta = (\pi_1, \pi_2, \ldots, \pi_K, t_1, t_2, \ldots, t_K).$$
In context of classification problems, a generative model is a statistical model \( \mathcal{P} \) on \( \mathcal{X} \times \{1, 2, \ldots, K\} \), where each \( P_\theta \in \mathcal{P} \) is

\[
P_\theta(x, y) = \pi_y \cdot P_{ty}(x),
\]

where the parameters are

\[
\theta = (\pi_1, \pi_2, \ldots, \pi_K, t_1, t_2, \ldots, t_K).
\]

Typically, all class conditional distributions \( P_{t_1}, P_{t_2}, \ldots, P_{t_K} \) come from same statistical model (e.g., Gaussian distribution family).
In context of classification problems, a *generative model* is a statistical model $\mathcal{P}$ on $\mathcal{X} \times \{1, 2, \ldots, K\}$, where each $P_\theta \in \mathcal{P}$ is

$$P_\theta(x, y) = \pi_y \cdot P_{t_y}(x),$$

where the parameters are

$$\theta = (\pi_1, \pi_2, \ldots, \pi_K, t_1, t_2, \ldots, t_K).$$

Typically, all class conditional distributions $P_{t_1}, P_{t_2}, \ldots, P_{t_K}$ come from same statistical model (e.g., Gaussian distribution family).

Form of Bayes classifier corresponding to $P_\theta$:

$$x \mapsto \arg \max_{y \in \{1, 2, \ldots, K\}} \pi_y \cdot P_{t_y}(x).$$
Learning a classifier

Basic approach to learning a classifier using a generative model
Learning a classifier

Basic approach to learning a classifier using a generative model

Suppose we observe data $D = \{(x_i, y_i)\}_{i=1}^n$, regarded as an i.i.d. sample.
Basic approach to learning a classifier using a generative model

Suppose we observe data $D = \{(x_i, y_i)\}_{i=1}^n$, regarded as an i.i.d. sample.

1. Split $D$ into $D_1, D_2, \ldots, D_K$, where

$$D_y = \{x_i : y_i = y\}.$$
Learning a classifier

Basic approach to learning a classifier using a generative model

Suppose we observe data $D = \{(x_i, y_i)\}_{i=1}^n$, regarded as an i.i.d. sample.

1. Split $D$ into $D_1, D_2, \ldots, D_K$, where

   $$D_y = \{x_i : y_i = y\}.$$

2. Estimate $\pi_1, \pi_2, \ldots, \pi_K$ using $\{y_i\}_{i=1}^n$ (e.g., using MLE: $\hat{\pi}_y := |D_y|/n$). For each $y \in \{1, 2, \ldots, K\}$: estimate $t_y$ using $D_y$ (e.g., using MLE).
**Learning a Classifier**

Basic approach to learning a classifier using a generative model

Suppose we observe data $D = \{(x_i, y_i)\}_{i=1}^n$, regarded as an i.i.d. sample.

1. Split $D$ into $D_1, D_2, \ldots, D_K$, where
   
   $$D_y = \{x_i : y_i = y\}.$$

2. Estimate $\pi_1, \pi_2, \ldots, \pi_K$ using $\{y_i\}_{i=1}^n$ (e.g., using MLE: $\hat{\pi}_y := |D_y|/n$).
   
   For each $y \in \{1, 2, \ldots, K\}$: estimate $t_y$ using $D_y$ (e.g., using MLE).

3. Classifier $\hat{f}$ is Bayes classifier for distribution $P_{\theta}$ corresponding to parameter estimates
   
   $$\hat{\theta} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_K, \hat{t}_1, \hat{t}_2, \ldots, \hat{t}_K),$$
   
   i.e.,
   
   $$\hat{f}(x) = \arg \max_{y \in \{1, 2, \ldots, K\}} \hat{\pi}_y \cdot P_{\hat{t}_y}(x).$$
**Example: Gaussian class conditional densities**

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, and using Gaussian class conditional densities.
Example: Gaussian class conditional densities

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, and using Gaussian class conditional densities.

Data $D = \{ (x_i, y_i) \}_{i=1}^n$, regarded as an i.i.d. sample.
Example: Gaussian class conditional densities

**Example**: \(\mathcal{X} = \mathbb{R}, \mathcal{Y} = \{0, 1\}\), and using Gaussian class conditional densities.

Data \(D = \{(x_i, y_i)\}_{i=1}^n\), regarded as an i.i.d. sample.

1. Split \(D\) into \(D_0, D_1\), where \(D_0 := \{x_i : y_i = 0\}\) and \(D_1 := D \setminus D_0\).
Example: Gaussian class conditional densities

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, and using Gaussian class conditional densities.

Data $D = \{(x_i, y_i)\}_{i=1}^n$, regarded as an i.i.d. sample.

1. Split $D$ into $D_0, D_1$, where $D_0 := \{x_i : y_i = 0\}$ and $D_1 := D \setminus D_0$.

2. Estimate parameters:

$$\hat{\pi}_0 := \frac{|D_0|}{n}, \quad \hat{\pi}_1 := \frac{|D_1|}{n},$$

$$\hat{\mu}_0 := \text{sample mean}(D_0), \quad \hat{\mu}_1 := \text{sample mean}(D_1),$$

$$\hat{\sigma}^2_0 := \text{sample variance}(D_0), \quad \hat{\sigma}^2_1 := \text{sample variance}(D_1).$$
Example: Gaussian class conditional densities

Example: $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, and using Gaussian class conditional densities.

Data $D = \{(x_i, y_i)\}_{i=1}^n$, regarded as an i.i.d. sample.

1. Split $D$ into $D_0, D_1$, where $D_0 := \{x_i : y_i = 0\}$ and $D_1 := D \setminus D_0$.

2. Estimate parameters:

$$\hat{\pi}_0 := |D_0|/n, \quad \hat{\pi}_1 := |D_1|/n,$$
$$\hat{\mu}_0 := \text{sample mean}(D_0), \quad \hat{\mu}_1 := \text{sample mean}(D_1),$$
$$\hat{\sigma}_0^2 := \text{sample variance}(D_0), \quad \hat{\sigma}_1^2 := \text{sample variance}(D_1).$$

3. Form classifier $\hat{f}$ using

$$\hat{f}(x) = \arg\max_{y \in \{0, 1\}} \hat{\pi}_y \cdot \varphi_{\hat{\mu}_y, \hat{\sigma}_y^2}(x)$$

where $\varphi_{\mu, \sigma^2}$ is the $N(\mu, \sigma^2)$ density.
Example: Gaussian class conditional densities

\[ \hat{\pi}_0 = \frac{1}{2}, \quad \hat{\mu}_0 = 0, \quad \hat{\sigma}_0^2 = 1 \]

\[ \hat{\pi}_1 = \frac{1}{2}, \quad \hat{\mu}_1 = 1, \quad \hat{\sigma}_1^2 = \frac{1}{4} \]

Classifier:

\[ \hat{f}(x) = \begin{cases} 1 & \text{if } x \in [0.38, 2.29]; \\ 0 & \text{otherwise.} \end{cases} \]

Dotted lines = decision boundary.
Suppose $\mathcal{X} = \{0, 1\}^d$ or $\mathcal{X} = \mathbb{Z}_+^d$ or $\mathcal{X} = \mathbb{R}^d$ where $d > 1$. 
Suppose $\mathcal{X} = \{0, 1\}^d$ or $\mathcal{X} = \mathbb{Z}_+^d$ or $\mathcal{X} = \mathbb{R}^d$ where $d > 1$.

What statistical model should we use for the class conditional distributions?
Suppose $\mathcal{X} = \{0, 1\}^d$ or $\mathcal{X} = \mathbb{Z}_+^d$ or $\mathcal{X} = \mathbb{R}^d$ where $d > 1$.

What statistical model should we use for the class conditional distributions?

- There are very general non-parametric models, but their quality may be poor when $d$ is large.
Suppose $\mathcal{X} = \{0, 1\}^d$ or $\mathcal{X} = \mathbb{Z}_+^d$ or $\mathcal{X} = \mathbb{R}^d$ where $d > 1$.

What statistical model should we use for the class conditional distributions?

- There are very general non-parametric models, but their quality may be poor when $d$ is large.

- Often you may have prior knowledge about the statistical dependencies between variables. Leverage this knowledge to form a graphical model.
Suppose $\mathcal{X} = \{0, 1\}^d$ or $\mathcal{X} = \mathbb{Z}_+^d$ or $\mathcal{X} = \mathbb{R}^d$ where $d > 1$.

What statistical model should we use for the class conditional distributions?

- There are very general non-parametric models, but their quality may be poor when $d$ is large.
- Often you may have prior knowledge about the statistical dependencies between variables. Leverage this knowledge to form a graphical model.
- Some simple models: multivariate Gaussians, product distributions.
Suppose $\mathcal{X} = \{0, 1\}^d$, and let $\mathcal{P}$ be all product distributions on $\{0, 1\}^d$. 
Suppose $\mathcal{X} = \{0, 1\}^d$, and let $\mathcal{P}$ be all product distributions on $\{0, 1\}^d$.

Parameters $\mu = (\mu_1, \mu_2, \ldots, \mu_d) \in [0, 1]^d$:

$$P_\mu(x_1, x_2, \ldots, x_d) = \prod_{j=1}^{d} \mu_j^{x_j} (1 - \mu_j)^{1-x_j}.$$
Product distributions on $\{0, 1\}^d$

Suppose $\mathcal{X} = \{0, 1\}^d$, and let $\mathcal{P}$ be all product distributions on $\{0, 1\}^d$.

Parameters $\mu = (\mu_1, \mu_2, \ldots, \mu_d) \in [0, 1]^d$:

$$P_\mu(x_1, x_2, \ldots, x_d) = \prod_{j=1}^{d} \mu_j^{x_j} (1 - \mu_j)^{1-x_j}.$$

If $(X_1, X_2, \ldots, X_d) \sim P_\mu$, then $X_1, X_2, \ldots, X_d$ are independent random variables, and

$$P_\mu(X_j = 1) = \mu_j.$$
Naïve Bayes classifiers

Generative models that use product distributions as class conditionals → Naïve Bayes classifiers.
Naïve Bayes classifiers

Generative models that use product distributions as class conditionals
→ Naïve Bayes classifiers.

(Using product distributions on \{0, 1\}^d as in previous slide.)
Form of Bayes classifier corresponding to distribution with parameters
\( \theta = (\pi_1, \pi_2, \ldots, \pi_K, \mu_1, \mu_2, \ldots, \mu_K) \):

\[
x \mapsto \arg \max_{y \in \{1,2,\ldots,K\}} \pi_y \cdot P_{\mu_y}(x)
\]
Naïve Bayes classifiers

Generative models that use product distributions as class conditionals → Naïve Bayes classifiers.

(Using product distributions on \( \{0, 1\}^d \) as in previous slide.)

Form of Bayes classifier corresponding to distribution with parameters \( \theta = (\pi_1, \pi_2, \ldots, \pi_K, \mu_1, \mu_2, \ldots, \mu_K) \):

\[
x \mapsto \arg \max_{y \in \{1, 2, \ldots, K\}} \pi_y \cdot P_{\mu_y}(x)
= \arg \max_{y \in \{1, 2, \ldots, K\}} \log \left( \pi_y \cdot P_{\mu_y}(x) \right)
\]
Naïve Bayes Classifiers

Generative models that use product distributions as class conditionals → Naïve Bayes classifiers.

(Using product distributions on \(\{0, 1\}^d\) as in previous slide.)

Form of Bayes classifier corresponding to distribution with parameters \(\theta = (\pi_1, \pi_2, \ldots, \pi_K, \mu_1, \mu_2, \ldots, \mu_K)\):

\[
x \mapsto \arg \max_{y \in \{1, 2, \ldots, K\}} \pi_y \cdot P_{\mu_y}(x)
= \arg \max_{y \in \{1, 2, \ldots, K\}} \log \left( \pi_y \cdot P_{\mu_y}(x) \right)
= \arg \max_{y \in \{1, 2, \ldots, K\}} \log \left( \pi_y \cdot \prod_{j=1}^{d} \mu_{y,j}^{x_j} (1 - \mu_{y,j})^{1-x_j} \right)
\]
Naïve Bayes classifiers

Generative models that use product distributions as class conditionals

Naïve Bayes classifiers.

(Using product distributions on \{0, 1\}^d as in previous slide.)

Form of Bayes classifier corresponding to distribution with parameters \(\theta = (\pi_1, \pi_2, \ldots, \pi_K, \mu_1, \mu_2, \ldots, \mu_K)\):

\[
\mathbf{x} \mapsto \arg \max_{y \in \{1, 2, \ldots, K\}} \pi_y \cdot P_{\mu_y}(\mathbf{x})
\]

\[
= \arg \max_{y \in \{1, 2, \ldots, K\}} \log \left( \pi_y \cdot P_{\mu_y}(\mathbf{x}) \right)
\]

\[
= \arg \max_{y \in \{1, 2, \ldots, K\}} \log \left( \pi_y \cdot \prod_{j=1}^{d} \mu_{y,j}^{x_j} (1 - \mu_{y,j})^{1-x_j} \right)
\]

\[
= \arg \max_{y \in \{1, 2, \ldots, K\}} \log \pi_y + \sum_{j=1}^{d} x_j \log \mu_{y,j} + (1 - x_j) \log(1 - \mu_{y,j})
\]
Naïve Bayes classifiers

Generative models that use product distributions as class conditionals

→ Naïve Bayes classifiers.

(Using product distributions on \(\{0, 1\}^d\) as in previous slide.)
Form of Bayes classifier corresponding to distribution with parameters
\(\theta = (\pi_1, \pi_2, \ldots, \pi_K, \mu_1, \mu_2, \ldots, \mu_K)\):

\[
x \mapsto \arg \max_{y \in \{1, 2, \ldots, K\}} \pi_y \cdot P_{\mu_y}(x)
\]

\[
= \arg \max_{y \in \{1, 2, \ldots, K\}} \log \left( \pi_y \cdot P_{\mu_y}(x) \right)
\]

\[
= \arg \max_{y \in \{1, 2, \ldots, K\}} \log \left( \pi_y \cdot \prod_{j=1}^{d} \mu_{y,j}^{x_j} (1 - \mu_{y,j})^{1-x_j} \right)
\]

\[
= \arg \max_{y \in \{1, 2, \ldots, K\}} \log \pi_y + \sum_{j=1}^{d} x_j \log \mu_{y,j} + (1 - x_j) \log(1 - \mu_{y,j})
\]

\[
= \arg \max_{y \in \{1, 2, \ldots, K\}} b_y + \langle w_y, x \rangle
\]

for some appropriate definition of \(b_y \in \mathbb{R}\) and \(w_y \in \mathbb{R}^d\) in terms of \(\pi_y\) and \(\mu_y\).
Multivariate Gaussians on $\mathbb{R}^d$
Multivariate Gaussian class conditionals

Example: \( \mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{0, 1\} \), and using multivariate Gaussian class conditional densities.
**Multivariate Gaussian class conditionals**

**Example**: \( \mathcal{X} = \mathbb{R}^d \), \( \mathcal{Y} = \{0, 1\} \), and using multivariate Gaussian class conditional densities.

Bayes classifier corresponding to distribution with parameters \( \theta = (\pi_0, \pi_1, \mu_0, \Sigma_0, \mu_1, \Sigma_1) \):

\[
\Sigma_0 = \Sigma_1
\]

Bayes classifier:
linear decision boundary

\[
\Sigma_0 \neq \Sigma_1
\]

Bayes classifier:
quadratic decision boundary
Sometimes, we must estimate (some of) the parameters of the class conditional distributions \textit{jointly}, rather than separately.
Sometimes, we must estimate (some of) the parameters of the class conditional distributions jointly, rather than separately.

Example: multivariate Gaussian class conditionals with shared covariance:

\[ t_1 = (\mu_1, \Sigma), \quad t_2 = (\mu_2, \Sigma), \quad \ldots, \quad t_K = (\mu_K, \Sigma). \]

This is called parameter tying.
Sometimes, we must estimate (some of) the parameters of the class conditional distributions \textit{jointly}, rather than separately.

\textbf{Example:} multivariate Gaussian class conditionals \textit{with shared covariance}:

\begin{align*}
t_1 &= (\mu_1, \Sigma), & t_2 &= (\mu_2, \Sigma), & \ldots, & t_K &= (\mu_K, \Sigma).
\end{align*}

This is called \textit{parameter tying}.

MLEs for $\pi_1, \pi_2, \ldots, \pi_K, \mu_1, \mu_2, \ldots, \mu_K$ given $\{(x_i, y_i)\}_{i=1}^n$:

\begin{align*}
\hat{\pi}_y &:= |D_y|/n, & \hat{\mu}_y &:= \text{sample mean}(D_y).
\end{align*}
Generative models with parameter tying

Sometimes, we must estimate (some of) the parameters of the class conditional distributions jointly, rather than separately.

Example: multivariate Gaussian class conditionals with shared covariance:

\[ t_1 = (\mu_1, \Sigma), \quad t_2 = (\mu_2, \Sigma), \quad \ldots, \quad t_K = (\mu_K, \Sigma). \]

This is called parameter tying.

MLEs for \( \pi_1, \pi_2, \ldots, \pi_K, \mu_1, \mu_2, \ldots, \mu_K \) given \( \{(x_i, y_i)\}_{i=1}^n \):

\[ \hat{\pi}_y := |D_y|/n, \quad \hat{\mu}_y := \text{sample mean}(D_y). \]

But what’s the MLE for the shared class conditional covariance \( \Sigma \)?
Sometimes, we must estimate (some of) the parameters of the class conditional distributions \textit{jointly}, rather than separately.

**Example:** multivariate Gaussian class conditionals \textit{with shared covariance}:

\[
t_1 = (\mu_1, \Sigma), \quad t_2 = (\mu_2, \Sigma), \quad \ldots, \quad t_K = (\mu_K, \Sigma).
\]

This is called \textit{parameter tying}.

MLEs for \(\pi_1, \pi_2, \ldots, \pi_K, \mu_1, \mu_2, \ldots, \mu_K\) given \(\{(x_i, y_i)\}_{i=1}^n\):

\[
\hat{\pi}_y := |D_y|/n, \quad \hat{\mu}_y := \text{sample mean}(D_y).
\]

But what’s the MLE for the shared class conditional covariance \(\Sigma\)?

\[
\hat{\Sigma} := \frac{1}{n} \sum_{y=1}^{K} \sum_{x_i \in D_y} (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^\top.
\]
Final remarks

Some redeeming qualities of classifiers based on generative models:

- Simple recipe, many variations.
- Can leverage domain knowledge about class conditional distributions.
- Can be very efficient when $K$ is large.

Critical drawbacks:

- Classifier relies on formula (via Bayes' rule) that assumes the estimated class priors and conditional distributions are perfect, which is not true.
- Modeling $P$ away from decision boundary between classes is wasted effort: not necessary for good classification.
Final remarks

Some redeeming qualities of classifiers based on generative models:

▶ Simple recipe, many variations.
▶ Can leverage domain knowledge about class conditional distributions.
▶ Can be very efficient when $K$ is large.

Critical drawbacks:

▶ Classifier relies on formula (via Bayes’ rule) that assumes the estimated class priors and conditional distributions are perfect, which is not true.
▶ Modeling $P$ away from decision boundary between classes is wasted effort: not necessary for good classification.