Homework 1 Part 1 - Solutions

Question 1 (5 points)

Let $\mathbf{X} = \mathbb{R}^2$ and consider the set of concepts of the form $c = \{(x, y) : x^2 + y^2 \leq r^2\}$ for some real number r. Show that this class can be (ϵ, δ)-PAC-learned from training data of size $n \geq \frac{1}{\epsilon} \ln \frac{1}{\delta}$.

(From Mohri et al. (2018) Foundations of Machine Learning, 2nd ed., MIT Press. Exercise 2.3.)

In here, we assume that the concept class is realizable, that is, there exists a circle that classifies the training samples with 0 error.

Consider the target concept c_r corresponding to the circle with radius r. Now choose an inner circle with the largest radius s that is smaller radius than r, c_s.

Assume that the probability of any point landing in the ring is $\geq \epsilon$. So, if a point lands on the inner circle with radius s, the probability of error is $> \epsilon$. This also means that the probability of a point to miss the ring is at most $1 - \epsilon$. Thus, for a sample of size n:

$$P(\text{error} > \epsilon) \leq (1 - \epsilon)^n \leq e^{-n\epsilon}$$

Setting δ to be greater than or equal to the right-hand side leads to

$$n \geq \frac{1}{\epsilon} \ln\left(\frac{1}{\delta}\right)$$

Question 2 (5 points)

Give a PAC-learning algorithm for the concept class $\mathcal{C}$ formed by closed intervals $[a, b]$ with $a, b \in \mathbb{R}$.

(From Mohri et al. (2018) Foundations of Machine Learning, 2nd ed., MIT Press. Exercise 2.8.)

Consider the hypothesis class to the a closed interval $I = [a, b]$. Now let I_S be the most specific hypothesis corresponding to the tightest closed interval containing samples from the positive class. If $P(I) < \epsilon$, since I_S is a smaller interval then $R(I_S) < \epsilon$. Let's assume that $P(I) \geq \epsilon$.

Now consider the two intervals I_L and I_R defined as:

$$\begin{aligned} I_L &= [a, x_l] \quad \text{with} \;\; x_l = \inf\{x : P([a, x]) \geq \epsilon/2\} \\ I_R &= [x_r, b] \quad \text{with} \;\; x_r = \inf\{x : P([x_r, b]) \geq \epsilon/2\} \end{aligned}$$

By this definition, if a point x lands in the interval $[a, x_l[$, the probability is less than or equal to $\epsilon/2$. Similarly, if a point x lands in the interval $]x_r, b]$, the probability is less than or equal to $\epsilon/2$. By the union bound, the probability of landing in either of those intervals is less than or equal than ϵ. Thus, if I_S overlaps both I_L and I_R, then its error region has probability at most ϵ. Thus, $R(I_S) > \epsilon$

implies that I_S does not overlap with either I_L or I_R, that is, either none of the training points falls in I_L or none falls in I_R. Thus, by the union bound,

$$\begin{aligned} P(R(I_S) > \epsilon) &\leq P(S \cap I_L \neq \emptyset) + P(S \cap I_R \neq \emptyset) \\ &\leq 2(1-\epsilon/2)^n \\ &\leq 2e^{-n\epsilon/2} \end{aligned}$$

Setting δ to match the right-hand side gives the sample complexity $n = \frac{2}{\epsilon}\ln\left(\frac{2}{\delta}\right)$ and proes the PAC-learning of closed intervals.

Question 3 (5 points)

Show that the VC dimension of the triangle hypothesis class is 7 in two dimensions. (Hint: For best separation, it is best to place the seven points equidistant on a circle.)

(From Alpaydin, Ethem. (2014) Introduction to Machine Learning, 3rd ed., MIT Press. Exercise 2.10.)

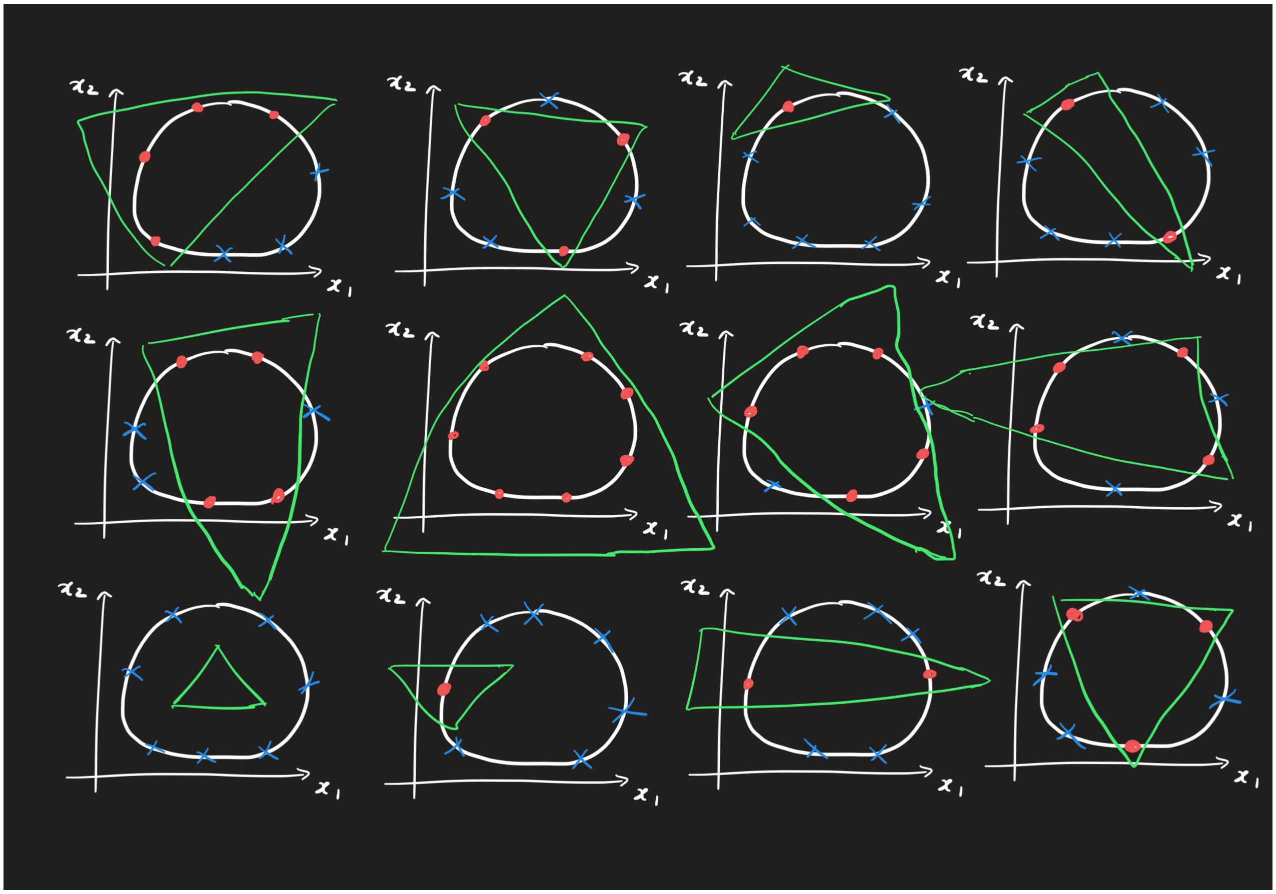
To show that the VC dimension of the triangle hypothesis class is 7 in two dimensions, you need to show that given a particular arrangement (e.g. points equidistant on a circle), all possible configurations/labeling of seven points, we can draw a triangle to separate the positive and negative examples.

Some examples include:

In [1]:

```
from IPython.display import Image
Image('figures/examples.jpg',width=700)
```

Out[1]:

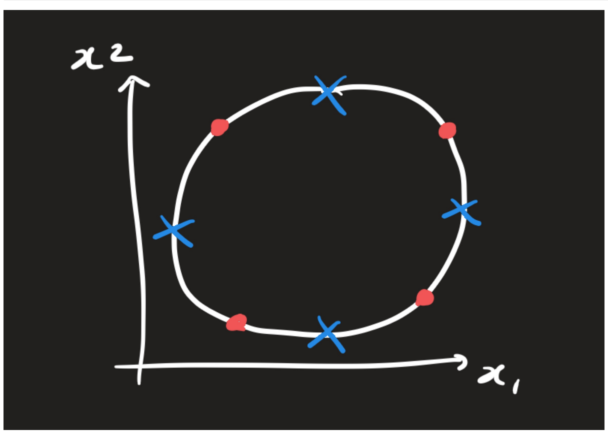


Moreover, we cannot do the same for 7+1=8 points, as the example below shows. (For any arrangement of 8 points, there is at least one configuration where the hypothesis does not shatter them).

In [2]:

```
Image('figures/triangle_class_on_8points.jpg',width=300)
```

Out[2]:



Question 4 (10 points)

Consider the Neyman-Pearson criteria for two univariate Cauchy distributions:

$$p(x|C_i) = \frac{1}{\pi\gamma}\left[1 + \left(\frac{x-\mu_i}{\gamma}\right)^2\right]^{-1} \qquad \forall i = 1, 2$$

where the mode is $\mu \in \mathbb{R}$ and width $\gamma \in \mathbb{R}_0^+$ ($\gamma > 0$).

Assume a zero-one error loss, and for simplicity $\mu_2 > \mu_1$, the same width γ, and equal prior.

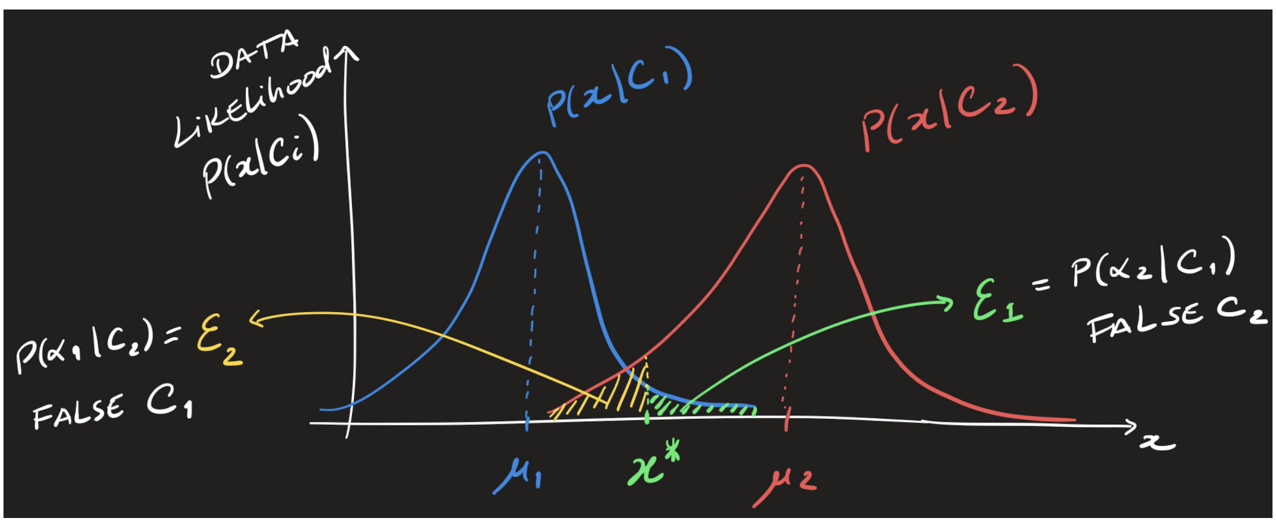
Answer the following questions:

1. **(2 points) Suppose the maximum acceptable error rate for classifying a pattern that is actually in C_1 as if it were in C_2 is ϵ_1. Determine the decision boundary in terms of the variables given.**
2. **(2 points) For this boundary, what is the error rate for classifying C_2 as C_1?**
3. **(2 points) What is the overall error rate under zero-one loss?**
4. **(2 points) Apply your results to the specific case $\gamma = 1$ and $\mu_1 = -1$, $\mu_2 = 1$ and $\epsilon_1 = 0.1$.**
5. **(2 points) Compare your result to the Bayes error rate (i.e., without the Neyman-Pearson conditions).**

(From Duda et al. (2001) Pattern Classification, 2nd ed., John Wiley & Sons. Exercise 2.6.)

In [3]:
```python
Image('figures/prob_of_error.png', width=700)
```

Out[3]:



1. As seen in the figure above, we have:

$$
\begin{aligned}
\epsilon_1 &= P(\alpha_2|C_1) \\
&= \int_{x^*}^{\infty} P(x|C_1)P(C_1)dx \\
&= \frac{1}{2}\int_{x^*}^{\infty} P(x|C_1)dx \\
&= \frac{1}{2}\int_{x^*}^{\infty} \frac{1}{\pi\gamma}\left[1+\left(\frac{x-\mu_1}{\gamma}\right)^2\right]^{-1} dx \\
&= \frac{1}{2\pi\gamma}\int_{x^*}^{\infty} \frac{1}{1+\left(\frac{x-\mu_1}{\gamma}\right)^2} dx
\end{aligned}
$$

Let $u = \frac{x-\mu_1}{\gamma} \Rightarrow x = u\gamma + \mu_1$ and $\frac{du}{dx} = \frac{1}{\gamma} \Rightarrow dx = \gamma du$.

$$\begin{aligned}\epsilon_1 &= \frac{1}{2\pi\gamma}\int_{x^*}^{\infty} \frac{1}{1+\left(\frac{x-\mu_1}{\gamma}\right)^2}dx \\ &= \frac{1}{2\pi\gamma}\int_{u\gamma+\mu_1}^{\infty} \frac{1}{1+u^2}\gamma du \\ &= \frac{1}{2\pi}\int_{u\gamma+\mu_1}^{\infty} \frac{1}{1+u^2}du\end{aligned}$$

Since $\sin\theta = \frac{1}{\sqrt{1+u^2}}$, letting $\sin(0) = \infty$ we get:

$$\begin{aligned}\epsilon_1 &= \frac{1}{2\pi}\int_{\hat{\theta}}^{0} d\theta \\ &= \frac{1}{2\pi}\sin^{-1}\left[\frac{\gamma}{\sqrt{\gamma^2+(x^*-\mu_1)^2}}\right]\end{aligned}$$

where $\hat{\theta} = \sin^{-1}\left[\frac{\gamma}{\sqrt{\gamma^2+(x^*-\mu_1)^2}}\right]$. Solving for the decision point x^* gives:

$$x^* = \mu_1 + \gamma\sqrt{\frac{1}{\sin^2(2\pi\epsilon_1)} - 1} = \mu_1 + \frac{\gamma}{\tan(2\pi\epsilon_1)}$$

1. As seen in the figure above, we have:

$$\begin{aligned}\epsilon_2 &= P(\alpha_1|C_2) \\ &= \int_{-\infty}^{x^*} P(x|C_2)P(C_2)dx\end{aligned}$$

Using a similar approach, we have:

$$\begin{aligned}\epsilon_2 &= \frac{1}{2\pi\gamma}\int_{-\infty}^{x^*} \frac{1}{1+\left(\frac{x-\mu_2}{\gamma}\right)^2}dx \\ &= \frac{1}{2\pi}\int_{-\pi}^{\hat{\theta}} d\theta \\ &= \frac{1}{2\pi}\left(\sin^{-1}\left[\frac{\gamma}{\sqrt{\gamma^2+(x^*-\mu_2)^2}}\right] + \pi\right) \\ &= \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\left[\frac{\gamma}{\sqrt{\gamma^2+(x^*-\mu_2)^2}}\right]\end{aligned}$$

1. Under the zero-one loss, both types of errors are equally likely with a loss of 1. Thus, the total error is:

$$\epsilon = \epsilon_1 + \epsilon_2 = \frac{1}{2\pi}\sin^{-1}\left[\frac{\gamma}{\sqrt{\gamma^2 + (x^* - \mu_1)^2}}\right] + \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\left[\frac{\gamma}{\sqrt{\gamma^2 + (x^* - \mu_2)^2}}\right]$$

where $x^* = \mu_1 + \frac{\gamma}{\tan(2\pi\epsilon_1)}$.

1. Letting $\gamma = 1$, $\mu_1 = -1$, $\mu_2 = 1$ and $\epsilon_1 = 0.1$, we have $x^* \approx 0.3764$ and find

$$\epsilon = 0.1 + \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\left[\frac{1}{\sqrt{1 + (0.3764 - 1)^2}}\right]$$

In [66]:

```
xp = -1 + 1/np.tan(2*np.pi*0.1)
xp
```

Out[66]:

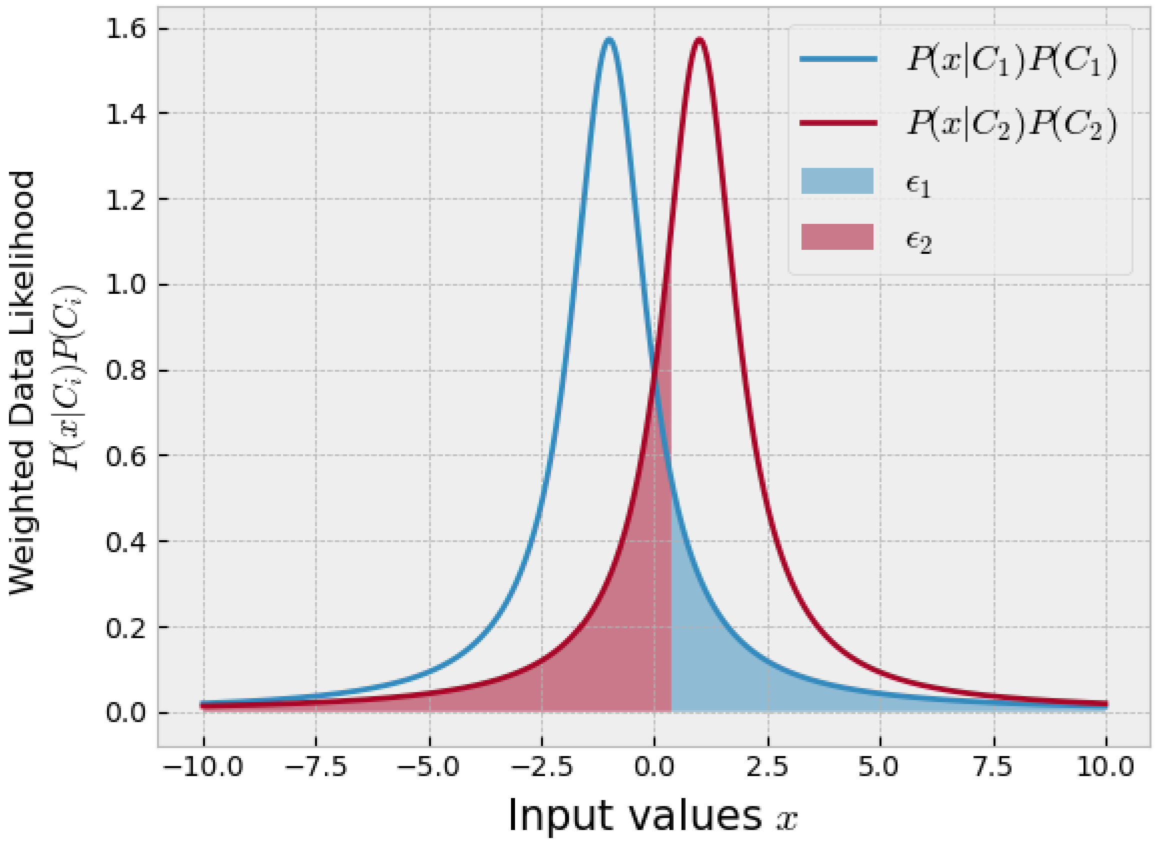
```
0.3763819204711736
```

In [63]:

```
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('bmh')

# Defining Cauchy likelihoods
mu1 = -1
mu2 = 1
gamma = 1
p1 = 0.5
p2 = 1 - p1
xplot=np.linspace(-10,10,1000)

cauchy=lambda x,mu,gamma: ((1/(np.pi*gamma))*(1+((x-mu)/gamma)**2))**(-1)
```

In [68]:

```
# Point (xp) in which epsilon_1 = 0.1

plt.plot(xplot, p1*cauchy(xplot, mu1, gamma), label='$P(x|C_1)P(C_1)$')
plt.plot(xplot, p2*cauchy(xplot, mu2, gamma), label='$P(x|C_2)P(C_2)$')
plt.fill_between(np.linspace(xp,10,100),
                 p1*cauchy(np.linspace(xp,10,100), mu1, gamma),
                 alpha=0.5, label='$\epsilon_1$')
plt.fill_between(np.linspace(-10,xp,100),
                 p2*cauchy(np.linspace(-10,xp,100), mu2, gamma),
                 alpha=0.5, label='$\epsilon_2$')

plt.xlabel('Input values $x$', size=15)
plt.ylabel('Weighted Data Likelihood \n$P(x|C_i)P(C_i)$')
plt.legend(fontsize=13);
```

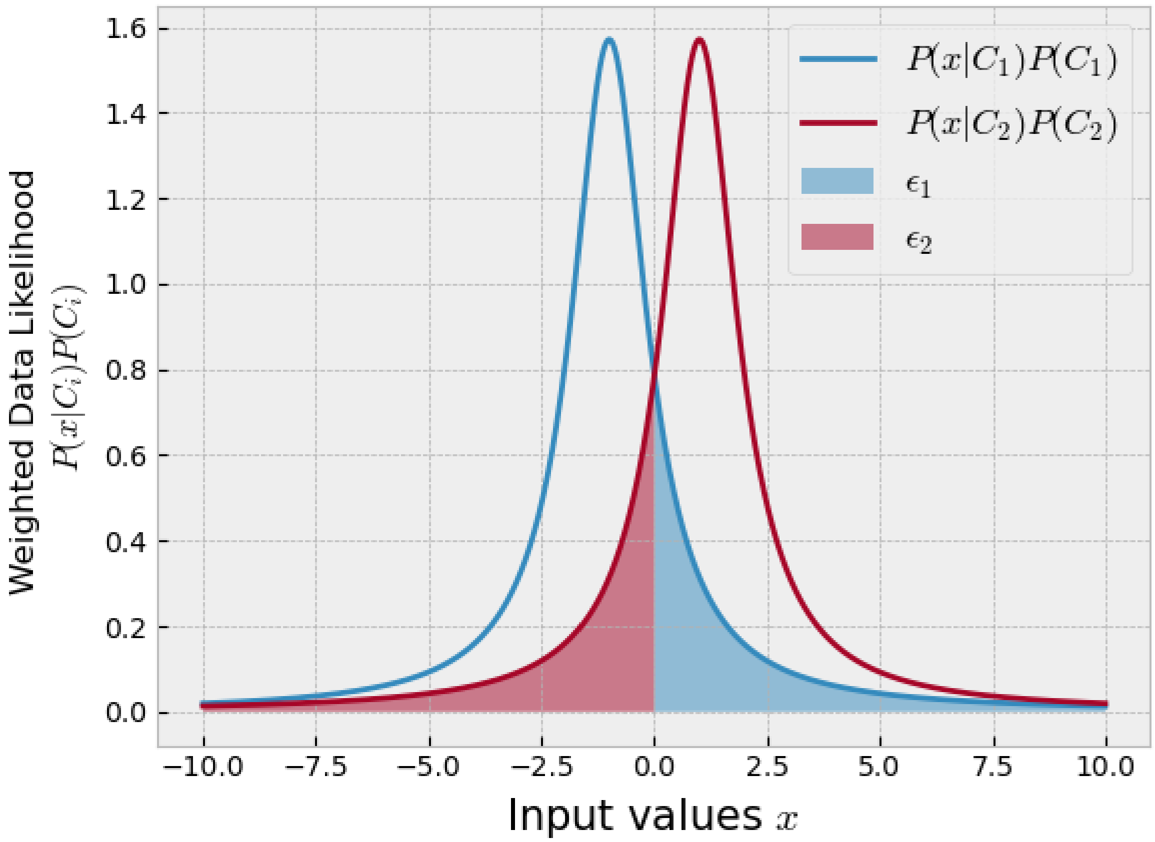


1. The Bayes' decision rule is the point in which the weighted data likelihood is largest, that is,

$$\text{Choose} \quad C_1 \quad \text{if} \quad \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)} \geq 0 \Rightarrow \frac{P(x|C_1)}{P(x|C_2)} \geq \frac{P(C_2)}{P(C_1)}$$

In [90]:

```
plt.plot(xplot, p1*cauchy(xplot, mu1, gamma), label='$P(x|C_1)P(C_1)$')
plt.plot(xplot, p2*cauchy(xplot, mu2, gamma), label='$P(x|C_2)P(C_2)$')
plt.fill_between(np.linspace(0,10,100),
                 p1*cauchy(np.linspace(0,10,100), mu1, gamma),
                 alpha=0.5, label='$\epsilon_1$')
plt.fill_between(np.linspace(-10,0,100),
                 p2*cauchy(np.linspace(-10,0,100), mu2, gamma),
                 alpha=0.5, label='$\epsilon_2$')

plt.xlabel('Input values $x$', size=15)
plt.ylabel('Weighted Data Likelihood \n$P(x|C_i)P(C_i)$')
plt.legend(fontsize=13);
```



As seen in the figure above, the midway point is $x = 0$. We can also calculate it:

In [91]:
```
xline=np.linspace(-10,10,100)

print('Midway point is: x=',
      round(xline[np.where(cauchy(xline, mu1, gamma)>
                          cauchy(xline, mu2, gamma))[0]][-1]))
```

```
Midway point is: x= 0
```

Since these distributions are symmetric, this corresponds to:

$$\begin{aligned}\epsilon_{\text{Bayes}} &= \int_{\infty}^{0} P(x|C_2)P(C_2)dx + \int_{0}^{\infty} P(x|C_1)P(C_1)dx \\ &= 2\int_{0}^{\infty} P(x|C_1)P(C_1)dx \\ &= 2\epsilon_1 \\ &= 2\frac{1}{2\pi}\sin^{-1}\left[\frac{\gamma}{\sqrt{\gamma^2 + (0-\mu_1)^2}}\right]\end{aligned}$$

where $\gamma = 1$ and $\mu_1 = -1$.

$$\epsilon_{\text{Bayes}} \approx 0.25$$

In [89]:
```
np.arcsin(gamma/np.sqrt(gamma**2 + (0-mu1)**2))/(np.pi)
```

Out[89]: 0.24999999999999997

Question 5 (8 points)

Consider two univariate Gaussian distributions:

$$p(x|C_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \qquad \forall i = 1, 2$$

where the mean is $\mu \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}$.

In addition, consider the following decision rule for this binary classification problem: _Decide C_1 if $x > \theta$; otherwise decide C_2_.

Answer the following questions:

1. **(2 points) Show that the probability of error for this rule is**

$$P(\text{error}) = P(C_1) \int_{-\infty}^{\theta} p(x|C_1)dx + P(C_2) \int_{\theta}^{\infty} p(x|C_2)dx.$$

1. **(2 points) By differentiating, show that a necessary condition to minimize $P(\text{error})$ is that θ satisfy**

$$P(\theta|C_1)P(C_1) = P(\theta|C_2)P(C_2).$$

1. **(2 points) Does this equation define θ uniquely?**

2. **(2 points) Give an example where a value of θ satisfying the equation actually maximizes the probability of error (i.e. specify values for μ_1, μ_2, σ and the priors).**

(From Duda et al. (2001) Pattern Classification, 2nd ed., John Wiley & Sons. Exercise 2.9.)

1. Under this decision rule, an error occurs when:
   - $x \in C_1$ but $x < \theta$, or
   - $x \in C_2$ but $x \geq \theta$.

Thus, we can write the probability of error as:

$$\begin{aligned}
P(\text{error}) &= \int_{-\infty}^{\infty} P(\text{error}|x)P(x)dx \\
&= P(x < \theta \cap x \in C_1) + P(x \geq \theta \cap x \in C_2) \\
&= P(x < \theta|C_1)P(C_1) + P(x \geq \theta|C_2)P(C_2) \\
&= \int_{-\infty}^{\theta} P(x|C_1)P(C_1)dx + \int_{\theta}^{\infty} P(x|C_2)P(C_2)dx \\
&= P(C_1) \int_{-\infty}^{\theta} P(x|C_1)dx + P(C_2) \int_{\theta}^{\infty} P(x|C_2)dx \blacksquare
\end{aligned}$$

1. First off, note that, since the data likelihoods are Gaussian distributed (and the fact they are symmetric w.r.t. the mean), we can rewrite the error as:

$$\begin{aligned}P(\text{error}) &= P(C_1)\int_{-\infty}^{\theta} P(x|C_1)dx + P(C_2)\int_{\theta}^{\infty} P(x|C_2)dx \\ &= P(C_1)\int_{-\infty}^{\theta} P(x|C_1)dx + P(C_2)\left(1-\int_{-\infty}^{\theta} P(x|C_2)dx\right) \\ &= P(C_1)\int_{-\infty}^{\theta} P(x|C_1)dx + P(C_2) - P(C_2)\int_{-\infty}^{\theta} P(x|C_2)dx\end{aligned}$$

Now, we can take the derivative of $P(\text{error})$ with respect to θ and set it to zero to find the critical points, that is

$$\frac{dP(\text{error})}{d\theta} = 0$$

Remember that, by the Fundamental Theorem of Calculus, for $F(x) = \int_a^b f(x)dx$, then $F'(x) = f(b) - f(a)$.

Coming back to the probability of error, we have:

$$\begin{aligned}P(C_1)(P(\theta|C_1) - P(-\infty|C_1)) - P(C_2)(P(\theta|C_2) - P(-\infty|C_2)) &= 0 \\ P(C_1)(P(\theta|C_1) - 0) - P(C_2)(P(\theta|C_2) - 0) &= 0 \\ P(C_1)P(\theta|C_1) &= P(C_2)P(\theta|C_2)\blacksquare\end{aligned}$$

1. No, it does not uniquely specify θ. In the Bayes' decision rule, for each choice of prior, we would find a different value for θ in which this condition is satisfied.

But also note that, if the distributions are unimodal and symmetric, like Gaussian distributions, for *a* given choice of priors $P(C_i), i = 1, 2$, there exists only one θ for which $P(C_1)P(\theta|C_1) = P(C_2)P(\theta|C_2)$. This is the midway point in Bayes' error.

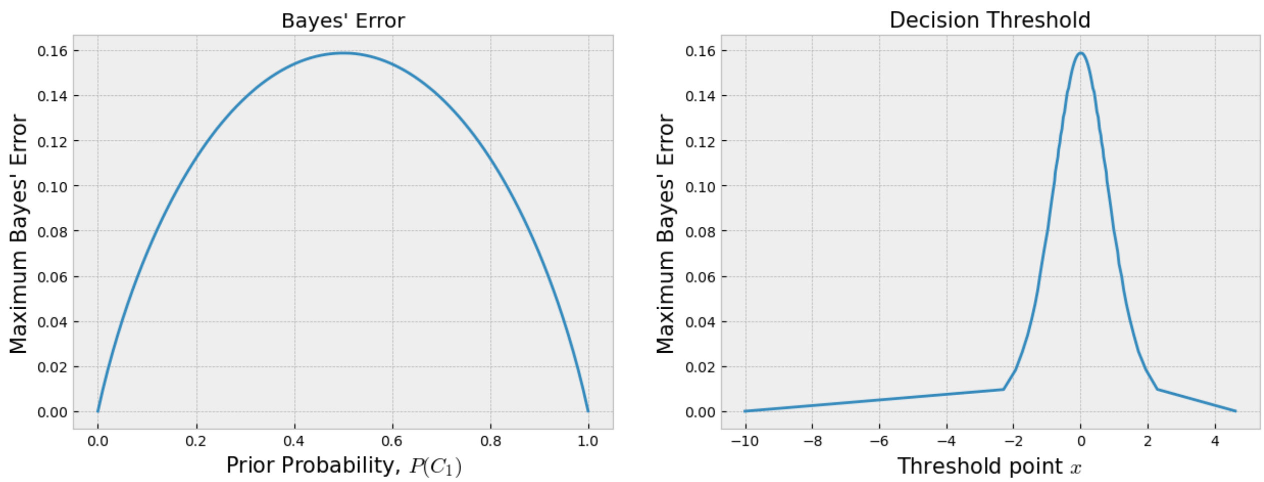
If the distributions are multimodal (e.g. mixture models), then, fixing the priors, this conditions may be satisfied for multiple values of θ.

1. Consider $\mu_1 = -1$, $\mu_2 = 1$ and $\sigma_1 = \sigma_2 = 1$, let's build a routine to find the largest error as a function of the prior probability $0 \leq P(C_1) \leq 1$.

In [114…

```
import scipy.stats as stats

G1 = stats.norm(loc=-1, scale=1)
G2 = stats.norm(loc=1,  scale=1)

xplot = np.linspace(-10,10,1000)

bayes_error = []
x_min = []
for p1 in np.linspace(0,0.9999,100):
```

```
    p2 = 1 - p1
    x = xplot[np.where(p2*G2.pdf(xplot)>p1*G1.pdf(xplot))[0]][0]
    x_min += [x]
    bayes_error += [p2*G2.cdf(x)+p1*G1.sf(x)]

plt.figure(figsize=(15,5))
plt.subplot(1,2,1); plt.plot(np.linspace(0,0.9999,100), bayes_error)
plt.xlabel('Prior Probability, $P(C_1)$', size=15)
plt.ylabel("Maximum Bayes' Error", size=15)
plt.title("Bayes' Error")
plt.subplot(1,2,2); plt.plot(x_min, bayes_error)
plt.xlabel('Threshold point $x$', size=15)
plt.ylabel("Maximum Bayes' Error", size=15)
plt.title('Decision Threshold', size=15);
```



We see that the maximum error, regardless of the prior probability, occurs at $x = 0$.

Question 6 (6 points)

Define action α_i as the decision to assign the input to class C_1 and λ_{ik} as the *loss* incurred for taking action α_i when the input actually belongs to C_k.

In a two-class, two-action problem, if the loss is $\lambda_{11} = \lambda_{22} = 0$, $\lambda_{12} = 10$, and $\lambda_{21} = 5$, write the optimal decision rule. How does the rule change if we add a third action of reject with $\lambda_{31} = \lambda_{32} = 1$?

(From Alpaydin, Ethem. (2014) Introduction to Machine Learning, 3rd ed., MIT Press. Exercise 3.4.)

The loss function table is as follows:

	α_1	α_2
C_1	$\lambda_{11} = 0$	$\lambda_{21} = 5$
C_2	$\lambda_{12} = 10$	$\lambda_{22} = 0$

The risk of action α_i is $R(\alpha_i|x) = \sum_{k=1}^{2} \lambda_{ik} P(C_k|x)$. Then, we find:

$$R(\alpha_1|x) = 10P(C_2|x) = 10(1 - P(C_1|x))$$
$$R(\alpha_2|x) = 5P(C_1|x)$$

Thus, the optimal (Bayesian/MAP) decision rule is to choose α_1 if:

$$\begin{aligned} R(\alpha_1|x) &< R(\alpha_2|x) \\ 10(1 - P(C_1|x)) &< 5P(C_1|x) \\ 10 - 10P(C_1|x) &< 5P(C_1|x) \\ P(C_1|x) &> \frac{2}{3} \end{aligned}$$

As expected, since the cost for choosing α_1 when it is actually C_2 is higher, we only choose α_1 when the posterior $P(C_1|x)$ is much larger than 1/2.

If we had a reject option with a cost of 1, the loss function table is as follows:

	α_1	α_2	α_r
C_1	$\lambda_{11} = 0$	$\lambda_{21} = 5$	$\lambda_{31} = \lambda = 1$
C_2	$\lambda_{12} = 10$	$\lambda_{22} = 0$	$\lambda_{32} = \lambda = 1$

The risk of each action is:

$$\begin{aligned} R(\alpha_1|x) &= 10P(C_2|x) \\ R(\alpha_2|x) &= 5P(C_1|x) \\ R(\alpha_r|x) &= 1 \end{aligned}$$

The optimal decision rule is:

$$\text{Choose } \ C_i \ \text{ if } \ R(\alpha_i|\mathbf{x}) < R(\alpha_k|\mathbf{x}) \ \text{ for all } k \neq i \text{ and } R(\alpha_i|\mathbf{x}) < \lambda$$
$$\text{Reject} \quad \text{otherwise}$$

Thus,

$$\begin{aligned} \text{Choose} \quad \alpha_1 &: \ 10P(C_2|x) < 1 \iff 10(1 - P(C_1|x)) < 1 \iff P(C_1|x) > \frac{9}{10} \\ \text{Choose} \quad \alpha_2 &: \ 5P(C_1|x) < 1 \iff P(C_1|x) < \frac{1}{5} \\ \text{Reject} \quad &: \ \frac{1}{5} < P(C_1|x) < \frac{9}{10} \end{aligned}$$

Question 7 (6 points)

An *association rule* is an implication of the form $X \to Y$ where X is the *antecedent* and Y is the *consequent* of the rule. One example of association rule is in *basket analysis* where we want to find the dependency between two items X and Y. In learning association rules, we can measure:

- ***Support* of the association rule $X \to Y$:**

$$\text{Support}(X, Y) \equiv P(X, Y) = \frac{\#\text{ customers who bought X and Y}}{\#\text{ customers}}$$

- ***Confidence* of the association rule $X \rightarrow Y$:**

$$\text{Confidence}(X, Y) \equiv P(Y|X) = \frac{P(X, Y)}{P(X)} = \frac{\# \text{ customers who bought X and Y}}{\# \text{ customers who bought X}}$$

Answer the following questions:

1. **(3 points) Given the following data of transactions at a shop, calculate the support and confidence values of** $\text{milk} \rightarrow \text{bananas}$**,** $\text{bananas} \rightarrow \text{milk}$**,** $\text{milk} \rightarrow \text{chocolate}$**, and** $\text{chocolate} \rightarrow \text{milk}$**.**

Transaction	Items in Basket
1	milk, bananas, chocolate
2	milk, chocolate
3	milk, bananas
4	chocolate
5	chocolate
6	milk, chocolate

1. **(3 points) Generalize the confidence and support formulas for basket analysis to calculate** k**-dependencies, namely** $P(Y|X_1, \ldots, X_k)$**.**

(From Alpaydin, Ethem. (2014) Introduction to Machine Learning, 3rd ed., MIT Press. Exercise 3.7.)

1. The association rules and their support and confidence values are as follows:

$$\begin{aligned}
\text{milk} \rightarrow \text{bananas} &\quad : \text{Support} = 1/6, \text{Confidence} = 2/4 \\
\text{bananas} \rightarrow \text{milk} &\quad : \text{Support} = 2/6, \text{Confidence} = 2/2 \\
\text{milk} \rightarrow \text{chocolate} &\quad : \text{Support} = 3/6, \text{Confidence} = 3/4 \\
\text{chocolate} \rightarrow \text{milk} &\quad : \text{Support} = 3/6, \text{Confidence} = 3/5
\end{aligned}$$

Though only half of the people who buy milk buy bananas too, anyone who buys bananas also buys milk.

1. We are interested in rules of the form $X_1, X_2, \ldots, X_k \rightarrow Y$:

   - Support:

$$P(X_1 \cap X_2 \cap \ldots X_k \cap Y) = \frac{\#\{\text{customers who bought } X_1 \text{ and } \ldots X_k \text{ and } Y\}}{\#\{\text{customers}\}}$$

   - Confidence:

$$P(Y|X_1 \cap X_2 \cap \cdots \cap X_k) = \frac{P(X_1 \cap X_2 \cap \cdots \cap X_k \cap Y)}{P(X_1 \cap X_2 \cap \cdots \cap X_k)} = \frac{\#\{\text{customers who bought}}{\#\{\text{customers who bou}}$$

Note that people who bought X_1, X_2, X_3 over a certain number should have bought X_1, X_2 and X_1, X_3 and X_2, X_3 over the same amount. So one can expand k-dependencies from $(k-1)$-dependencies. This is the basic idea behind the **Apriori algorithm**.

On-Time (5 points)

Submit your assignment before the deadline.
