**ESE 504-542 : Statistics for Data Science**
**Instructor: Hamed Hassani**
**Fall 2019**

# Final Examination

| NAME | |
|---|---|

One two-sided note-sheet allowed.

| | Grade (y/n) | Score | Max. Score |
|---|---|---|---|
| Problem 1 | | | 40 |
| Problem 2 | | | 30 |
| Problem 3 | | | 30 |
| TOTAL | | | 100 |

**Problem 1 (40 points)**

Recall that in classification we assume that each data point is an i.i.d. sample from a distribution $P(X = x, Y = y)$. In this question, we are going to consider a specific data distribution $P$ and evaluate the performance of logistic regression and Bayes optimal classifier on data generated using $P$. In the following, we assume $x \in \mathbb{R}$ and $y \in \{-1, 1\}$, i.e. the data is one-dimensional and the label is binary. Write $P(X = x, Y = y) = P(Y = y)P(X = x|Y = y)$. We let $P(y = +1) = P(Y = -1) = \frac{1}{2}$ and

$$P(X = x|Y = +1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-5)^2}{2}), \text{ and,}$$

$$P(X = x|Y = -1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x+5)^2}{2}).$$

1. Start from $P(X = x, Y = y) = P(Y = y)P(X = x|Y = y)$ and show that $P(X = x, Y = y) = \frac{1}{2\sqrt{2\pi}} \exp(-\frac{(x-5y)^2}{2})$. (This is a simple one line derivation.)

We can re-state given equations into:

$$P(X=x \mid Y=y) = \frac{1}{\sqrt{2\pi}} exp(-\frac{(X-5y)^2}{2})$$

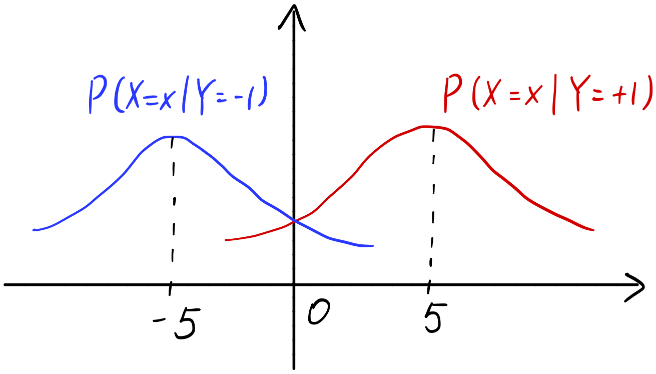
$$\therefore \quad P(X=x, Y=y)$$

$$= P(Y=y)\ P(X=x|Y=y)$$

$$= \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} exp(-\frac{(X-5y)^2}{2})$$

$$= \frac{1}{2\sqrt{2\pi}} exp(-\frac{(X-5y)^2}{2})$$

2. Plot the conditional distributions $P(X = x|Y = +1)$ and $P(X = x|Y = -1)$ in one figure. I.e. you should plot two gaussian pdfs in one figure.



3. Write the Bayes optimal classification rule given the above distribution $P$ and simplify it (hint: in the end you should reach to a very simple classification rule that classifies an input $x$ based on whether or not its value is greater than a threshold).

$$P(Y=y \mid X=x) = \frac{P(X=x, Y=y)}{P(X=x)}$$

$$= P(Y=y \mid X=x) \cdot \frac{P(Y=y)}{P(X=x)}$$

$$\therefore \hat{y} = h(x) = \operatorname*{argmax}_{y} \; P(Y=y \mid X=x)$$

$$= \operatorname*{argmax}_{y} \; P(Y=y \mid X=x) \cdot \frac{P(Y=y)}{P(X=x)}$$

$$= \operatorname*{argmax}_{y} \; P(X=x \mid Y=y)$$

$$= \begin{cases} +1 & , \text{ if } \; x > 0 \\ -1 & , \text{ otherwise} \end{cases} \qquad \text{by plot}$$

4. Compute the probability of classification error for the Bayes optimal classifier.

error

$$= \mathbb{E}_{(X,Y)\sim P}\left[ \mathbb{1}\left(h(x) \neq Y\right) \right]$$

$$= P_{(X,Y)\sim P}\left(h(x) \neq Y\right)$$

$$= P(X>0)\, P(Y=-1 \mid X=x) + P(X<0)\, P(Y=+1 \mid X<0)$$

$$= P(Y=-1, X>0) + P(Y=+1, X<0)$$

$$= P(Y=-1)\, P(X>0 \mid Y=-1) + P(Y=+1)\, P(X<0 \mid Y=+1)$$

$$= 2 \cdot \frac{1}{2} \left(1 - \Phi\left(\frac{5}{1}\right)\right)$$

$$= 1 - \Phi(5)$$

5. Let us now consider logistic regression (this part and the next can be answered independently from the previous parts). Given training data $(x_1, y_1), \cdots, (x_n, y_n)$, explain briefly the main steps of training a logistic regression model. I.e. what quantities/probabilities are being estimated by logistic regression? What is the parametric model used? How are the parameters of the model optimized?

Estimate: $P(Y=+1|X=x)$, $P(Y=-1|X=x)$

Model: $$P(Y=+1|X=x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

$$P(Y=-1|X=x) = \frac{1}{1+e^{\beta_0+\beta_1 x}}$$

Optimization:

$$L(\beta_0, \beta_1 | x) = Pr(Y | X ; \beta_0, \beta_1)$$
$$= \prod_{i=1}^{n} P(Y=y_i | X=x_i ; \beta_0, \beta_1)$$

$$\max_{\beta_0, \beta_1} L(\beta_0, \beta_1 | x)$$

6. Going back to the data distribution $P$ detailed above, logistic regression needs to find the value of two parameters $\beta_0$ and $\beta_1$ using training data $\{(x_i, y_i)\}_{i=1,\cdots,n}$ generated according to the distribution $P$. Assume that the number of training data points is very large (i.e. $n \to \infty$); What will be the parameters $\beta_0$ and $\beta_1$ in this case? (Hint: Start by deriving the exact form of the conditional distribution $P(Y = y|X = x)$.)

From: $P(X=x \mid Y=+1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{2}\right), P(Y=+1) = \frac{1}{2}$

$P(X=x \mid Y=-1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x+5)^2}{2}\right), P(Y=-1) = \frac{1}{2}$

$$
\begin{aligned}
& P(Y=+1 \mid X=x) \\
= & \frac{P(X=x, Y=+1)}{P(X=x, Y=+1) + P(X=x, Y=-1)} \\
= & \frac{P(X=x \mid Y=+1) P(Y=+1)}{P(X=x \mid Y=+1) P(Y=+1) + P(X=x \mid Y=-1) P(Y=-1)} \\
= & \frac{\frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{2}\right)}{\frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-5)^2}{2}\right) + \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x+5)^2}{2}\right)} \\
= & \frac{\exp\left(-\frac{(x-5)^2}{2}\right)}{\exp\left(-\frac{(x-5)^2}{2}\right) + \exp\left(-\frac{(x+5)^2}{2}\right)} \\
= & \frac{1}{\exp\left(\frac{(x-5)^2}{2} - \frac{(x+5)^2}{2}\right) + 1} \\
= & \frac{1}{\exp(-10x) + 1}
\end{aligned}
$$

Note that logistic regression would try estimate

$P(Y=+1 \mid X=x) = \frac{1}{1+\exp(-\beta_1 x - \beta_0)}$

When $n$ is large, which means there is enough data to estimate probability accurately,

So $\beta_0 = 0$, $\beta_1 = 10$.

**Problem 2 (30 points)** [**Weighted K-Means Clustering**.]
Consider data points $x_1, x_2, \cdots, x_n \in \mathbb{R}^p$. We aim to provide an algorithm for the following clustering problem: Find $K$ centers $c_1, c_2, \cdots, c_K \in \mathbb{R}^p$ that minimize the objective

$$\sum_{i=1}^{n} \min_{j \in \{1, \cdots, K\}} ||x_i - c_j||_1, \tag{1}$$

where $|| \cdot ||_1$ is the $L_1$ (Manhattan) distance.

1. Assume that $K = 1$. Find the optimal centroid that minimizes (1).

$K = 1$

$$f(c) = \sum_{i=1}^{n} \| X_i - c \|_1$$

$$= \sum_{i=1}^{n} \sum_{\ell=1}^{p} | X_{i\ell} - c_\ell |$$

$$\frac{\partial f}{\partial c} = \sum_{i=1}^{n} \sum_{\ell=1}^{p} \text{sign}( c_\ell - X_{i\ell} ) = 0$$

$$\therefore \ |\{ x_i | \ X_{i\ell} > c_\ell \}| = |\{ x_i | X_{i\ell} < c_\ell \}|$$

$\therefore$ $c_\ell$ is the median of

$X_{1\ell}, X_{2\ell} \cdots X_{n\ell}$ , for all $1 \leq \ell \leq p$

2. For a given $K$, derive an iterative algorithm to find a good set of centers for the above problem. Explain precisely what the algorithm is and justify your answer. (Hint: Recall the steps of the K-Means algorithm done in class and see how you could change those for the above setting).

1. Randomly initialize K centers

2. Assign data points to cluster with nearest (least Manhattan distance) center.

3. Re-calculate centers with $C_{jl}$ to be the median of $\{X_{jl} \mid X_j \in C_j\}$, for all $1 \leq l \leq p$

4. Repeat 2, 3 until centers don't update.

**Problem 3 (30 points)** **[Short answer questions]**

1. In which cases should we prefer LDA to QDA and vice versa? Briefly justify your answer.

LDA requires less parameters, and
is more likely to underfit and
less likely to overfit than QDA.
So we prefer LDA to QDA
when dataset is small
(we don't want overfitting)
and vice versa when dataset is large
(we don't want underfitting)

2. Assume that we have a data matrix $X$ of dimension $p \times n$ as usual. Suppose that its SVD is of the from $X = USV^T$, where $S$ is a diagonal matrix with $s_1 = 10$ and $s_2 = s_3 = \cdots = s_p = 1$. Assume that we want to compress the data from $p$ to 1 dimensions via a linear transform represented by a $1 \times p$ matrix $C$ and then reconstruct via $p \times 1$ matrix $R$. Let $\tilde{X} = RCX$ be the reconstruction. What is the smallest reconstruction error that can be achieved?

Compress data from $p$ to 1 dimension :

$U = [u_1, \cdots, u_p]$, $C = u_1^T$

We choose the first principal component with $R = u_1$, since the largest eigenvalue leads to the smallest reconstruction error.

$\| X - \tilde{X} \|_2^2 = \| X - RCX \|_2^2 = \| USV^T - US^{(1)}V^T \|_2^2$

$S^{(1)}_{ij} = 10 \cdot 1 \, (i=1, j=1)$

$\therefore \| X - \tilde{X} \|_2^2 = \| U(S - S^{(1)}) V^T \|_2^2$

$= \| S - S^{(1)} \|_2^2$

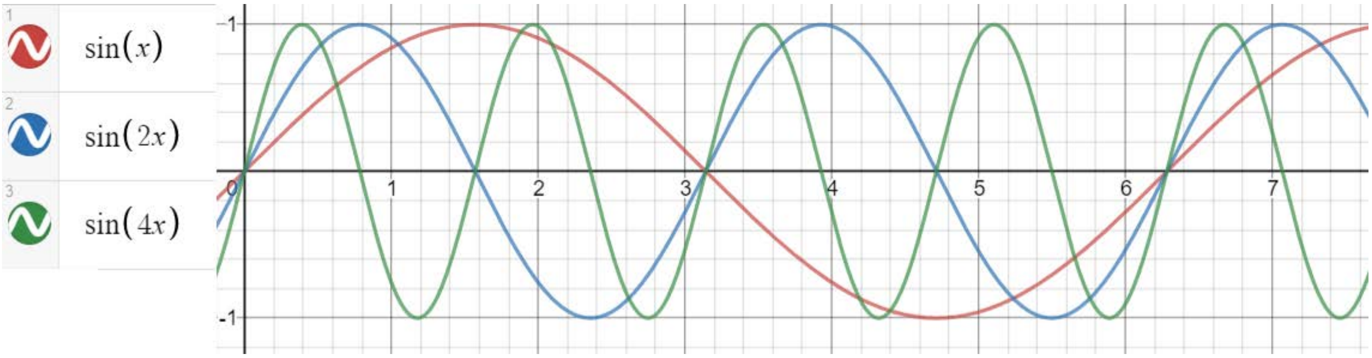
$= \sum_{i=2}^{p} s_i^2$

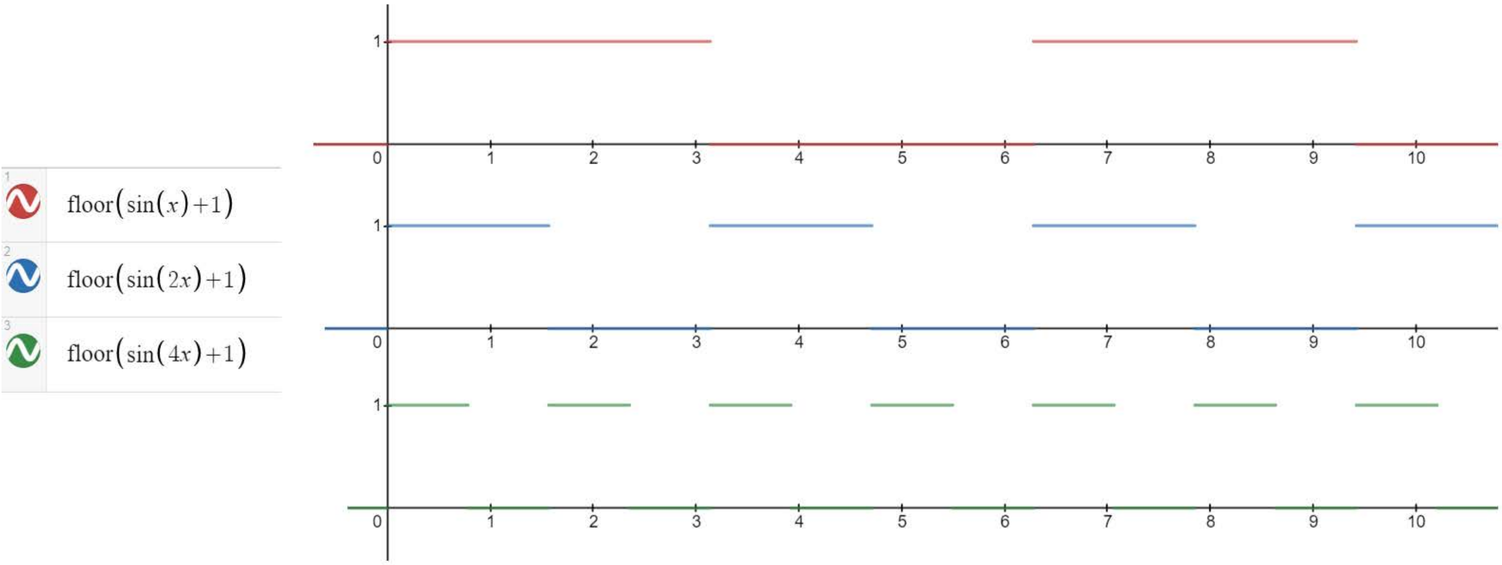
$= p - 1$

( If we choose other principle components,
$\| X - \tilde{X} \|_2^2 = p + 8 > p - 1$ )

3. Consider the hypothesis class $\mathcal{H} = \{h_\theta(x) = \lceil \sin(\theta x) \rceil, \theta \in \mathbb{R}\}$. What is the VC dimension of $\mathcal{H}$? Recall that $\lceil y \rceil$ is the smallest integer that's larger than $y$. (hint: Try to plot a few functions from $\mathcal{H}$ for different values of $\theta$.)

Intuitively, when $\theta$ increases, $\sin(\theta x)$ become more and more complex. Below is illustration:





So $\mathcal{H}$ should shatter sets of any cardinality, VC-dim $= \infty$.

And following (not required) is a way to construct such set $C$.

**ESE 504-542 : Statistics for Data Science**
**Instructor: Hamed Hassani, Shirin Saeedi**
**Spring 2019**

# Final Examination

| | |
|---|---|
| NAME | |

One two-sided note-sheet allowed.

| | Grade (y/n) | Score | Max. Score |
|---|---|---|---|
| Problem 1 | | | 40 |
| Problem 2 | | | 40 |
| Problem 3 | | | 20 |
| TOTAL | | | 100 |

**Problem 1 (40 points)** [**Simple Linear Regression.**]
Consider the following simple linear regression problem with the data set $(x_1, y_1)$, $(x_2, y_2), \ldots,$ $(x_n, y_n)$.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

Assume all assumptions for linear regression are met. Particularly, $\epsilon_i$ are i.i.d. random variables where $\epsilon_i \sim N(0, \sigma^2)$.

1. Derive the estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ by minimizing the residual sum of squares i.e., by solving

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

2. Derive the estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ using maximum likelihood estimation i.e., by solving

$$\max_{\beta_0,\beta_1} \ \log \ell(\beta_0, \beta_1),$$

where

$$\ell(\beta_0, \beta_1) = \prod_{i=1}^{n} \Pr\left(y_i \mid \beta_0, \beta_1, x_i\right).$$

Note that $\Pr\left(y_i \mid \beta_0, \beta_1, x_i\right)$ is the probability of observing $y_i$ given the values of $\beta_0, \beta_1$ and $x_i$. Compare the results with part 1.

3. Show that your estimates are unbiased i.e., show that

$$E\left[\hat{\beta}_0\right] = \beta_0, \qquad E\left[\hat{\beta}_1\right] = \beta_1.$$

4. Consider the case when heteroskedasticity is present, i.e., $\epsilon_i \sim N(0, \sigma_i^2)$. Repeat part 2 under heteroskedasticity.

**Problem 2 (40 points)** [**Weighted K-Means Clustering**.]

Consider data points $x_1, x_2, \cdots, x_n \in \mathbb{R}^d$. For each data point $x_i$ we have assigned a positive number $w_i \geq 0$ which indicates the importance of that data point. Our goal is to provide an algorithm for the following *weighted K-Means clustering* problem: Find $K$ centers $c_1, c_2, \cdots, c_K \in \mathbb{R}^d$ that minimize the objective

$$\sum_{i=1}^{n} w_i \times \min_{j \in \{1, \cdots, K\}} ||x_i - c_j||^2. \tag{2}$$

1. Assume that $K = 1$. Find the optimal centroid that minimizes (2).

2. For a given $K$, Extend the K-Means algorithm taught in class to the weighted setting. Explain precisely what the algorithm is and justify your answer

**Problem 3 (20 points)** [**Basic Questions about Learning Theory.**]

1. Give a precise definition of “PAC Learnability”.

2. Explain briefly why finite hypothesis classes are PAC learnable.

3. What property should an infinite hypothesis class have in order to be PAC learnable?

**ESE 402-542 : Statistics for Data Science**
**Instructor: Hamed Hassani**
**Fall 2020**

# Final Examination

| NAME | |
|---|---|

| | Grade (y/n) | Score | Max. Score |
|---|---|---|---|
| Problem 1 | | | 30 |
| Problem 2 | | | 40 |
| Problem 3 | | | 30 |
| TOTAL | | | 100 |

**Problem 1 (30 points)**
Assume that $X_1, X_2, \cdots, X_n$ are generated i.i.d. according to the following distribution:

$$\Pr(X = i) = \theta(1-\theta)^i, \text{ for } i = 0, 1, 2, 3, \cdots$$

In other words, the pdf (pmf) of the distribution that generates the data is of the form $f(X = i|\theta) = \theta(1-\theta)^i$, where $\theta$ is an unknown parameter. Consider the following hypothesis testing problem:

$$\begin{aligned} H_0: &\quad \theta = \theta_0 \\ H_{\mathrm{a}}: &\quad \theta = \theta_a, \end{aligned}$$

where we assume that $\theta_a > \theta_0$.

Derive the most powerful test for this hypothesis testing problem and specify what the acceptance/rejection regions are for a given significance level $\alpha_0$ (you can assume that $n$ is large).

**Problem 2 (40 points)**

In this question we assume that data is generated according to a distribution $P(X = x, Y = y)$ given as follows: $x \in \mathbb{R}$ and $y \in \{-1, 1\}$, i.e. the data is one-dimensional and the label is binary. Write $P(X = x, Y = y) = P(Y = y)P(X = x|Y = y)$. We let $P(y = +1) = \frac{3}{4}$, and $P(Y = -1) = \frac{1}{4}$, and

$$P(X = x|Y = +1) = \frac{1}{2}\exp(-|x-2|), \text{ and,}$$

$$P(X = x|Y = -1) = \frac{1}{2}\exp(-|x+2|).$$

I.e. $P(X = x|Y = +1)$ and $P(X = x|Y = -1)$ follow the Laplace distribution. (The mean of a Laplace distribution with pdf $\frac{1}{2b}e^{(-\frac{|x-\mu|}{b})}$ is $\mu$ and the variance is $2b^2$.)

1. Derive the expression for $P(X = x, Y = y)$.

2. Plot the conditional distributions $P(X = x, Y = +1)$ and $P(X = x, Y = -1)$ in one figure.

3. Write the Bayes optimal classification rule given the above distribution $P$ and simplify it (hint: in the end you should reach to a very simple classification rule that classifies an input $x$ based on whether or not its value is greater than a threshold).

4. Compute the probability of classification error for the Bayes optimal classifier.

5. Let us now consider QDA (this part and the next can be answered independently from the previous parts). Given training data $(x_1, y_1), \cdots, (x_n, y_n)$, explain briefly the main steps of training the QDA model. I.e. what quantities/probabilities are being estimated by QDA? What is the parametric model used? How are the parameters of the model found? (Recall that QDA is similar to LDA with the exception that the variance per class can be different.)

6. Assume that the number of training data points is very large (i.e. $n \to \infty$); What will be the exact value of the parameters of the trained QDA model in this case? (Hint: You don't really need much calculation to derive the parameters.)

7. Simplify the QDA classifier that you obtained in the previous part (hint: in the end you should reach to a very simple classification rule that classifies an input $x$ based on whether or not its value is greater than a threshold).

8. Given your answers in Parts 2 and 5, what do you think about the performance of QDA compared to what can be done optimally? Does QDA perform optimally when we have many training data points?

**Problem 3 (30 points)**

1. Find the VC-dimension of the function class $\mathcal{H}_1 = \{h_a; a \in\mathbb{R}\}$, where $h_a : \mathbb{R} \to \{0, 1\}$ is defined as

$$h_a(x) = \begin{cases} 1, & \text{if } x \in [a, a+1] \\ 1, & \text{if } x \in [a+5, a+6] \\ 0 & \text{otherwise.} \end{cases}$$

2. Find the VC-dimension of the function class $\mathcal{H}_2 = \{h_{a,b}; a, b \in[0, +\infty)\}$, where $h_{a,b} : \mathbb{R} \to \{0, 1\}$ is defined as

$$h_{a,b}(x) = \begin{cases} 1, & \text{if } |x| \leq a \\ 1 & \text{if } |x| \geq b \\ 0 & \text{otherwise.} \end{cases}$$

3. Consider the class $\mathcal{H}_1$. Given a training data set $S$, let $h_S^*$ be the outcome of the empirical risk minimization procedure restricted to class $\mathcal{H}_1$. How many training samples do we need (ignoring constants) to ensure that for any distribution of the data, $\mathcal{D}$, the following event occurs with probability $1-\delta$?

$$L_{\mathcal{D}}(h_S^*) - \min_{h\in\mathcal{H}_1} L_{\mathcal{D}}(h) \leq \epsilon.$$

(Justify your answer in one sentence.)

4. Consider again the same class $\mathcal{H}_1$. Recall that we defined $h_S^*$ to be the outcome of the empirical risk minimization procedure restricted to class $\mathcal{H}_1$. How many training samples do we need (ignoring constants) to ensure that for any distribution of the data, $\mathcal{D}$, the following event occurs with probability $1-\delta$?

$$L_{\mathcal{D}}(h_S^*) - \min_{h\in\{\text{all the possible functions}\}} L_{\mathcal{D}}(h) \leq \epsilon.$$

(Justify your answer in one sentence.)