Question

A two-class classifier uses Bayesian decision theory. Given that the prior probabilities of two classes are $P(C_1) = 0.6$ and $P(C_2) = 0.4$, and their likelihoods follow normal distributions with means $\mu_1 = 5$, $\mu_2 = 8$ and equal variance $\sigma^2 = 4$, determine the decision boundary.

Answer

In Bayesian decision theory, we classify an observation x to the class C_k that maximizes the posterior probability $P(C_k \mid x)$. The decision boundary between two classes C_1 and C_2 is the point (or region) where the posterior probabilities are equal:

$$P(C_1 \mid x) = P(C_2 \mid x)$$

Using Bayes' Theorem, $P(C_k \mid x) = \frac{P(x|C_k)P(C_k)}{P(x)}$. Substituting this into the boundary condition:

$$\frac{P(x \mid C_1)P(C_1)}{P(x)} = \frac{P(x \mid C_2)P(C_2)}{P(x)}$$

Since $P(x)$ is the same for both classes and positive, we can simplify to:

$$P(x \mid C_1)P(C_1) = P(x \mid C_2)P(C_2)$$

This can be rewritten as a likelihood ratio:

$$\frac{P(x \mid C_1)}{P(x \mid C_2)} = \frac{P(C_2)}{P(C_1)}$$

We are given the prior probabilities $P(C_1) = 0.6$ and $P(C_2) = 0.4$.

$$\frac{P(C_2)}{P(C_1)} = \frac{0.4}{0.6} = \frac{4}{6} = \frac{2}{3}$$

The likelihoods are 1D normal distributions $\mathcal{N}(x \mid \mu_k, \sigma^2)$, with $\mu_1 = 5$, $\mu_2 = 8$, and $\sigma^2 = 4$. The probability density function is $P(x \mid C_k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)$.

Substitute the likelihoods and priors into the boundary equation $\frac{P(x|C_1)}{P(x|C_2)} = \frac{P(C_2)}{P(C_1)}$:

$$\frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right)} = \frac{P(C_2)}{P(C_1)}$$

The $\frac{1}{\sqrt{2\pi\sigma^2}}$ terms cancel out:

$$\frac{\exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(x-\mu_2)^2}{2\sigma^2}\right)} = \frac{P(C_2)}{P(C_1)}$$

$$\exp\left(\frac{(x-\mu_2)^2-(x-\mu_1)^2}{2\sigma^2}\right) = \frac{P(C_2)}{P(C_1)}$$

Take the natural logarithm of both sides:

$$\frac{(x-\mu_2)^2-(x-\mu_1)^2}{2\sigma^2} = \ln\left(\frac{P(C_2)}{P(C_1)}\right)$$

Substitute the given values $\mu_1 = 5, \mu_2 = 8, \sigma^2 = 4$, and $\frac{P(C_2)}{P(C_1)} = \frac{2}{3}$:

$$\frac{(x-8)^2-(x-5)^2}{2(4)} = \ln\left(\frac{2}{3}\right)$$

$$\frac{(x-8)^2-(x-5)^2}{8} = \ln\left(\frac{2}{3}\right)$$

Expand the squares in the numerator:

$$(x^2-16x+64)-(x^2-10x+25) = 8\ln\left(\frac{2}{3}\right)$$

$$x^2-16x+64-x^2+10x-25 = 8\ln\left(\frac{2}{3}\right)$$

Simplify the left side:

$$-6x+39 = 8\ln\left(\frac{2}{3}\right)$$

Now, solve for x to find the decision boundary:

$$-6x = 8\ln\left(\frac{2}{3}\right) - 39$$

$$6x = 39 - 8\ln\left(\frac{2}{3}\right)$$

$$x = \frac{39-8\ln\left(\frac{2}{3}\right)}{6}$$

This is the exact value of the decision boundary.

To get an approximate numerical value, $\ln(2/3) \approx -0.4055$:

$$x \approx \frac{39 - 8(-0.4055)}{6} = \frac{39 + 3.244}{6} = \frac{42.244}{6} \approx 7.04$$

The decision boundary is at $x = \frac{39 - 8\ln(2/3)}{6}$.

Based on the inequality $P(C_1 \mid x) > P(C_2 \mid x)$, which simplifies to $x < \frac{39 - 8\ln(2/3)}{6}$ for this specific setup (since C_1 has smaller mean and higher prior), the decision rule is:

- Classify as C_1 if $x < \frac{39 - 8\ln(2/3)}{6}$
- Classify as C_2 if $x > \frac{39 - 8\ln(2/3)}{6}$
- The boundary point is $x = \frac{39 - 8\ln(2/3)}{6}$.

Question

Maximum Likelihood Estimation (MLE) is commonly used in pattern recognition. Derive the MLE formula for a Gaussian distribution with unknown mean and known variance. Apply it to estimate the mean given the data points: {3, 7, 5, 9, 6}.

Answer

To derive the Maximum Likelihood Estimation (MLE) formula for the mean (μ) of a Gaussian distribution $\mathcal{N}(x \mid \mu, \sigma^2)$ when the variance (σ^2) is known, we need to find the value of μ that maximizes the likelihood of observing the given data.

Let the observed data points be $x = \{x_1, x_2, \ldots, x_n\}$, which are assumed to be independent and identically distributed (i.i.d.) samples from $\mathcal{N}(\mu, \sigma^2)$.

1. Write the Likelihood Function

The likelihood function $L(\mu \mid x)$ is the joint probability of observing the data x given the parameter μ. Since the samples are i.i.d., this is the product of the individual probability density functions (PDFs):

$$L(\mu \mid x) = \prod_{i=1}^{n} P(x_i \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

2. Write the Log-Likelihood Function

Maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood (log-likelihood), which is often simpler mathematically:

$$\ell(\mu \mid x) = \log L(\mu \mid x) = \log \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right) \right)$$

Using the property $\log(ab) = \log a + \log b$ and $\log(a^k) = k \log a$:

$$\ell(\mu \mid x) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \sum_{i=1}^{n} \log\left(\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right) \right)$$

$$\ell(\mu \mid x) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma^2) \right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Since $\log(2\pi\sigma^2)$ is constant with respect to μ:

$$\ell(\mu \mid x) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

3. Maximize the Log-Likelihood (Find the Derivative)

To find the value of μ that maximizes $\ell(\mu \mid x)$, we take the derivative with respect to μ and set it to zero. The first term is constant with respect to μ.

$$\frac{d\ell(\mu \mid x)}{d\mu} = \frac{d}{d\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \right)$$

$$\frac{d\ell(\mu \mid x)}{d\mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{d}{d\mu}(x_i-\mu)^2$$

Using the chain rule ($\frac{d}{d\mu}(x_i-\mu)^2 = 2(x_i-\mu)\cdot(-1) = -2(x_i-\mu)$):

$$\frac{d\ell(\mu \mid x)}{d\mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2(x_i-\mu))$$

$$\frac{d\ell(\mu \mid x)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

4. Set the Derivative to Zero and Solve for μ

Set the derivative equal to zero to find the critical point(s):

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

Since σ^2 is a known positive value, we can multiply by σ^2:

$$\sum_{i=1}^{n}(x_i-\mu) = 0$$

Distribute the sum:

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu = 0$$

The sum of μ for n terms is $n\mu$:

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

Solve for μ:

$$n\mu = \sum_{i=1}^{n} x_i$$

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The MLE estimate for the mean of a Gaussian distribution (with known variance) is the **sample mean** of the data.

Apply to Estimate the Mean for Data {3, 7, 5, 9, 6}

We have the data points $\{3, 7, 5, 9, 6\}$.
The number of data points is $n = 5$.
The sum of the data points is $\sum x_i = 3 + 7 + 5 + 9 + 6 = 30$.

Using the derived MLE formula for the mean ($\hat{\mu}_{MLE} = \frac{1}{n} \sum x_i$):

$$\hat{\mu}_{MLE} = \frac{1}{5} \times 30$$

$$\hat{\mu}_{MLE} = 6$$

The Maximum Likelihood Estimate for the mean of the Gaussian distribution, given the data points {3, 7, 5, 9, 6} and known variance, is 6.

Question

You are building an automatic disease diagnosis system that uses Bayesian classification to distinguish between "disease-positive" and "disease-negative" patients. Given:

- $P(\text{Positive}) = 0.3$, $P(\text{Negative}) = 0.7$
- Likelihoods: $P(\text{Symptom} \mid \text{Positive}) = 0.8$, $P(\text{Symptom} \mid \text{Negative}) = 0.2$

Compute the posterior probability $P(\text{Positive} \mid \text{Symptom})$ using Bayes' theorem. Interpret the result in the context of medical diagnosis.

Answer

We can compute the posterior probability $P(\text{Positive} \mid \text{Symptom})$ using Bayes' theorem:

$$P(\text{Positive} \mid \text{Symptom}) = \frac{P(\text{Symptom} \mid \text{Positive})P(\text{Positive})}{P(\text{Symptom})}$$

We are given $P(\text{Symptom} \mid \text{Positive}) = 0.8$ (the likelihood of the symptom given the patient is positive) and $P(\text{Positive}) = 0.3$ (the prior probability of being positive).

To use Bayes' theorem, we first need to calculate the overall probability of observing the symptom, $P(\text{Symptom})$. We can do this using the law of total probability, considering the two possible states (Positive or Negative):

$$P(\text{Symptom}) = P(\text{Symptom} \mid \text{Positive})P(\text{Positive}) + P(\text{Symptom} \mid \text{Negative})P(\text{Negative})$$

We are given $P(\text{Symptom} \mid \text{Negative}) = 0.2$ and $P(\text{Negative}) = 0.7$.

$$P(\text{Symptom}) = (0.8)(0.3) + (0.2)(0.7)$$

$$P(\text{Symptom}) = 0.24 + 0.14$$

$$P(\text{Symptom}) = 0.38$$

The overall probability of observing the symptom in the population is 0.38.

Now we can compute the posterior probability $P(\text{Positive} \mid \text{Symptom})$:

$$P(\text{Positive} \mid \text{Symptom}) = \frac{P(\text{Symptom} \mid \text{Positive})P(\text{Positive})}{P(\text{Symptom})} = \frac{(0.8)(0.3)}{0.38}$$

$$P(\text{Positive} \mid \text{Symptom}) = \frac{0.24}{0.38} = \frac{24}{38} = \frac{12}{19}$$

As a decimal, $\frac{12}{19} \approx 0.6316$.

The posterior probability $P(\text{Positive} \mid \text{Symptom}) \approx 0.6316$.

Interpretation in the context of medical diagnosis:

- The **prior probability ($P(\text{Positive}) = 0.3$)** represents the initial belief or prevalence of the disease in the population *before* considering any symptoms. In this case, 30% of the patients are expected to be disease-positive based on general population data.
- The **likelihoods ($P(\text{Symptom} \mid \text{Positive}) = 0.8$ and $P(\text{Symptom} \mid \text{Negative}) = 0.2$)** tell us how informative the symptom is. The symptom is significantly more likely to occur in patients who are truly disease-positive (80%) than in those who are disease-negative (20%).
- The **posterior probability ($P(\text{Positive} \mid \text{Symptom}) \approx 0.6316$)** is the updated probability of a patient being disease-positive *after* observing the specific symptom.

The result shows that observing this symptom more than doubles the probability of a patient being disease-positive, increasing it from a prior of 30% to approximately 63.16%. This symptom is a valuable indicator for diagnosing the disease. Based on this single symptom, a Bayesian classifier would likely classify a patient exhibiting this symptom as "disease-positive" (assuming a standard threshold like 0.5 for the posterior probability).

Question

Suppose you have a dataset with 1000 samples and 100 features per sample. If the number of features is increased to 500, discuss how this affects classifier performance and suggest techniques to mitigate issues arising due to the "curse of dimensionality."

Answer

Increasing the number of features from 100 to 500 while keeping the number of samples at 1000 significantly impacts classifier performance due to the phenomenon known as the **"curse of dimensionality."**

How Increasing Features Affects Classifier Performance:

1. **Data Sparsity:** In a high-dimensional space (500 features), the available 1000 data points become extremely sparse. The volume of the space grows exponentially with the number of dimensions, so the fixed number of samples occupies a rapidly decreasing fraction of the space. Data points that might seem close in lower dimensions become very far apart in higher dimensions.
2. **Meaningless Distances:** As dimensionality increases, the distance between any two points tends to become less discriminative. The ratio between the nearest and farthest points approaches 1, making distance-based methods (like k-Nearest Neighbors or clustering) less effective as points are all "far" from each other.
3. **Increased Risk of Overfitting:** With a high number of features relative to the number of samples (500 features vs. 1000 samples – a ratio of 1:2), the classifier can easily find complex patterns or correlations in the training data that are just noise or random chance. The model fits the training data perfectly but fails to generalize to new, unseen data. This is a major problem of **overfitting**.
4. **Increased Variance:** The variance of model estimates increases with dimensionality because there is less data available per dimension to accurately estimate parameters.
5. **Computational Cost:** Training and testing classifiers take significantly longer and require more memory as the number of features increases. Many algorithms' complexity scales

poorly with dimensionality.

6. **Difficulty in Visualization and Interpretation:** It becomes impossible to visualize data directly in 500 dimensions, making exploratory data analysis and interpretation of model behavior much harder.

In summary, going from 100 to 500 features with only 1000 samples makes the learning problem much harder. The dataset becomes sparse, models are prone to overfitting, and computational demands increase.

Techniques to Mitigate Issues Arising Due to the "Curse of Dimensionality":

To address these problems, especially the sparsity and overfitting issues, you can employ dimensionality reduction techniques or use models robust to high dimensions and overfitting:

1. **Dimensionality Reduction:** Reduce the number of features before training the classifier.
   - **Feature Selection:** Choose a subset of the *original* features that are most relevant to the prediction task.
     - *Filter Methods:* Rank features based on statistical measures (e.g., correlation with the target, mutual information) and select the top-ranked.
     - *Wrapper Methods:* Use a classifier to evaluate subsets of features (e.g., recursive feature elimination, forward/backward selection). Computationally expensive.
     - *Embedded Methods:* Feature selection is built into the model training process (e.g., L1 regularization / Lasso).
   - **Feature Extraction:** Create a smaller set of *new* features that are combinations or transformations of the original features while retaining most of the important information.
     - **Principal Component Analysis (PCA):** Unsupervised technique that finds orthogonal directions (principal components) that capture the maximum variance in the data. You can select the top few components that explain a sufficient amount of variance.
     - **Linear Discriminant Analysis (LDA):** Supervised technique that finds directions that maximize class separability (if you have class labels).
     - *Manifold Learning:* Non-linear methods like t-SNE or UMAP (often used for visualization) or Autoencoders (can be used for non-linear dimensionality reduction).
2. **Regularization:** Use techniques during model training to penalize model complexity.
   - **L1 (Lasso) or L2 (Ridge) Regularization:** Add a penalty term to the loss function of linear models, logistic regression, or neural networks that discourages large weights. L1 can shrink some weights to exactly zero, effectively performing feature selection.
   - **Dropout:** In neural networks, randomly deactivate neurons during training to prevent co-adaptation and encourage robustness.
3. **Choose Appropriate Models:** Some models perform better in high dimensions than others.

    - Linear models with strong regularization (as mentioned above).
    - Tree-based ensemble methods (like Random Forests or Gradient Boosting Machines) are often less sensitive to irrelevant features than linear models, but they can still suffer from sparsity and computational cost.
4. **Gather More Data:** The most effective, albeit often impractical, solution is to increase the number of training samples. The required number of samples for effective learning grows exponentially with dimensionality.

For your specific case (1000 samples, 500 features), **Dimensionality Reduction** (both feature selection and extraction like PCA) and **Regularization** would be the primary techniques to consider. You would likely want to reduce the number of features significantly (perhaps back down towards or even below 100, depending on the data's intrinsic dimensionality and signal-to-noise ratio) before applying many standard classifiers effectively.

Question

Consider a face recognition system that classifies images into "known" and "unknown" categories. The system uses a Gaussian Mixture Model (GMM) to represent the distribution of facial features. Explain how the decision boundary is formed and how the Bayesian approach improves classification accuracy

---

Answer

Okay, let's explain the decision boundary and the role of the Bayesian approach in this system simply.

Think of a **GMM** for a class (like "Known") as a way to model the "shape" of the typical feature vectors for that class using a mix of several simple Gaussian bumps. The GMM tells you how likely a given feature vector x is if it came from that class ($P(x \mid \text{Class})$).

How the Decision Boundary is Formed

The system needs to decide if a new face's features (x) belong to the "Known" class or the "Unknown" class. Bayesian decision theory helps make this decision by comparing the **likelihood of the data given the class** multiplied by the **prior probability of the class**.

The rule is basically: Classify as "Known" if $P(x \mid \text{Known})P(\text{Known})$ is greater than $P(x \mid \text{Unknown})P(\text{Unknown})$.

The **decision boundary** is simply the line (or surface in high dimensions) where these two values are equal:

$$P(x \mid \text{Known})P(\text{Known}) = P(x \mid \text{Unknown})P(\text{Unknown})$$

Since the GMMs ($P(x \mid \text{Class})$) can have complex, bumpy shapes (because they are mixtures of Gaussians), the decision boundary formed by balancing the "Known" GMM (weighted by its prior) against the "Unknown" GMM (weighted by its prior) can be **curvy or non-linear**, not just a simple straight line.

How the Bayesian Approach (using Priors) Improves Accuracy

The "Bayesian approach" here mainly refers to using the **prior probabilities** ($P(\text{Known})$ and $P(\text{Unknown})$).

- **Priors** represent your initial belief about how likely a face is to be "known" or "unknown" *before* you even look at its specific features. For example, if this is a security system, you might set $P(\text{Unknown})$ very high because most people trying to enter might be unknown.
- By including the priors in the decision rule, you're not just asking "Which GMM shape does this feature vector *fit* best?". You're asking "Which class is more likely, considering both how well the features fit its model *AND* how probable that class is overall?".
- This makes the decision **smarter and more robust**. If "Unknown" has a much higher prior, the system will require very strong evidence from the GMM likelihood ($P(x \mid \text{Known})$ being much higher than $P(x \mid \text{Unknown})$) to classify a face as "Known". This helps to **reduce false positives** (classifying an unknown person as known), which is often a critical goal in security.

In short, using the Bayesian approach means using priors to bias the decision towards the more probable class, leading to better accuracy by reducing costly errors based on real-world frequencies.

Question

A fair six-sided die is rolled twice. Let X be the sum of the two outcomes.
• (a) Find the probability mass function (PMF) of X. (5 Marks)
• (b) Compute P(X ≥ 8).

Answer

Let the outcomes of the two fair six-sided dice rolls be D_1 and D_2. The possible outcomes for a single fair die are $\{1, 2, 3, 4, 5, 6\}$, each with a probability of $\frac{1}{6}$. When rolling two dice independently, there are $6 \times 6 = 36$ equally likely outcomes in the sample space, represented as ordered pairs (D_1, D_2).

Let $X = D_1 + D_2$ be the sum of the two outcomes. The possible values for X range from $1 + 1 = 2$ to $6 + 6 = 12$.

(a) Find the probability mass function (PMF) of X.

The PMF, $P(X = k)$, is the probability that the sum of the two dice is equal to a specific value k. This is calculated by counting the number of combinations of (D_1, D_2) that sum to k and dividing by the total number of outcomes (36).

Here are the outcomes for each possible sum k:

- $X = 2$: (1, 1) - 1 outcome
- $X = 3$: (1, 2), (2, 1) - 2 outcomes
- $X = 4$: (1, 3), (2, 2), (3, 1) - 3 outcomes
- $X = 5$: (1, 4), (2, 3), (3, 2), (4, 1) - 4 outcomes
- $X = 6$: (1, 5), (2, 4), (3, 3), (4, 2), (5, 1) - 5 outcomes
- $X = 7$: (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1) - 6 outcomes
- $X = 8$: (2, 6), (3, 5), (4, 4), (5, 3), (6, 2) - 5 outcomes
- $X = 9$: (3, 6), (4, 5), (5, 4), (6, 3) - 4 outcomes
- $X = 10$: (4, 6), (5, 5), (6, 4) - 3 outcomes
- $X = 11$: (5, 6), (6, 5) - 2 outcomes
- $X = 12$: (6, 6) - 1 outcome

The PMF of X is:

k	Outcomes (count)	$P(X = k)$
2	1	$1/36$
3	2	$2/36 = 1/18$
4	3	$3/36 = 1/12$
5	4	$4/36 = 1/9$
6	5	$5/36$
7	6	$6/36 = 1/6$
8	5	$5/36$
9	4	$4/36 = 1/9$
10	3	$3/36 = 1/12$
11	2	$2/36 = 1/18$
12	1	$1/36$

(b) Compute P(X ≥ 8).

$P(X \geq 8)$ is the probability that the sum X is 8 or more. This includes the probabilities $P(X = 8), P(X = 9), P(X = 10), P(X = 11)$, and $P(X = 12)$.

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$$

Using the probabilities from the PMF:

$$P(X \geq 8) = \frac{5}{36} + \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36}$$

$$P(X \geq 8) = \frac{5 + 4 + 3 + 2 + 1}{36} = \frac{15}{36}$$

This fraction can be simplified by dividing the numerator and denominator by their greatest common divisor, which is 3:

$$P(X \geq 8) = \frac{15 \div 3}{36 \div 3} = \frac{5}{12}$$

The probability $P(X \geq 8)$ is $\frac{15}{36}$ or $\frac{5}{12}$.

Question

The heights of students in a university are normally distributed with a mean of 170 cm and a standard deviation of 8 cm.
- (a) What is the probability that a randomly chosen student is taller than 180 cm?
- (b) What height represents the 90th percentile of this distribution?

Answer

Let H be the height of a randomly chosen student. H is normally distributed with a mean $\mu = 170$ cm and a standard deviation $\sigma = 8$ cm. We can write this as $H \sim \mathcal{N}(170, 8^2)$.

(a) Probability that a randomly chosen student is taller than 180 cm

We want to find $P(H > 180)$. To do this, we convert the height of 180 cm to a standard Z-score. The Z-score measures how many standard deviations a value is from the mean:

$$Z = \frac{H - \mu}{\sigma}$$

For $H = 180$ cm:

$$Z = \frac{180 - 170}{8} = \frac{10}{8} = 1.25$$

So, $P(H > 180)$ is equivalent to $P(Z > 1.25)$, where Z is the standard normal random variable ($\mathcal{N}(0, 1)$).

Using a standard normal (Z) table or calculator, we find the probability $P(Z \leq 1.25)$. This value is approximately 0.8944.
The probability $P(Z > 1.25)$ is $1 - P(Z \leq 1.25)$:

$$P(Z > 1.25) = 1 - 0.8944 = 0.1056$$

The probability that a randomly chosen student is taller than 180 cm is approximately 0.1056, or 10.56%.

(b) Height representing the 90th percentile

The 90th percentile is the height h_{90} such that 90% of students are shorter than or equal to that height. In terms of probability, this means we want to find h_{90} such that $P(H \leq h_{90}) = 0.90$.

First, we find the Z-score (z_{90}) corresponding to a cumulative probability of 0.90. We look up 0.90 in the body of a standard normal (Z) table and find the corresponding Z-score on the margins. The Z-score that corresponds to a cumulative probability of 0.90 is approximately 1.28. (Using a more precise value like 1.2816 is also common).

Now, we convert this Z-score back into a height using the Z-score formula, rearranged to solve for H:

$$h_{90} = \mu + z_{90}\sigma$$

Substitute the values $\mu = 170$, $z_{90} \approx 1.28$, and $\sigma = 8$:

$$h_{90} \approx 170 + (1.28)(8)$$

$$h_{90} \approx 170 + 10.24$$

$$h_{90} \approx 180.24 \text{ cm}$$

The height that represents the 90th percentile of this distribution is approximately 180.24 cm. This means 90% of students are shorter than or equal to 180.24 cm.