Solutions to Chapter 12

Prob. 12.1 — Classification corresponds to which branch of machine learning?

Answer (Prob. 12.1) — Classification is almost synonymous with supervised learning.

Prob. 12.2 — List the three types of classification problems, and briefly describe each.

Answer (Prob. 12.2) — The three types of classification problems are detection (search an image to find all instances of a particular type of object), recognition (determine the identity of an object), and verification (make a binary decision about whether the hypothesized identity of an object in correct).

Prob. 12.3 — Explain the importance of generalization.

Answer (Prob. 12.3) — Generalization refers to how well a classifier will perform on data that it has never seen. A classifier with poor generalization my perform well on a test set, but it will not be useful to anyone, because its performance on new data is bad. On the other hand, a classifier with good generalization performs well not only on the test set but also on new data.

Prob. 12.4 — Define a dichotomizer.

Answer (Prob. 12.4) — A dichotomizer is a classifier that distinguishes between exactly two categories.

Prob. 12.5 — Suppose an inspection system detects 99.99% of the good parts correctly, but only 98% of the defective parts. Suppose the cost of incorrectly labeling a good part bad is 5 minutes of extra time for a person to manually inspect the part, but the cost of incorrectly labeling a bad part good is an average of 3 hours of extra time to correct the problem downstream in the assembly line. What is the total risk of the system? Which type of error dominates the total risk?

Answer (Prob. 12.5) — The total risk is

$$\begin{aligned}\text{total risk} &= p(\text{label is bad}|\text{part is good}) \cdot 5 + p(\text{label is good}|\text{part is bad}) \cdot 180 \\ &= 0.01(5) + 2(180) \\ &= 360.05 \text{ min.}\end{aligned}$$

The total risk is dominated by incorrectly labeling the parts good when they are actually defective.

Prob. 12.6 — You are part of a team whose job is to develop a classifier of some kind. So you collect some data and manually label them. What are your two options regarding how to separate the training data from the test data? What are the pros and cons of each? What are you absolutely not allowed to do in any circumstance with these two datasets?

Answer (Prob. 12.6) — One option (holdout method) is to divide the data into two parts: training data and test data. Another option (e.g., LOOCV) is to iteratively leave out some of the data for testing the classifier that results from training on the rest of the data. The former yields accuracy numbers that are indicative of an actual classifier that will be used at runtime, whereas the latter allows the system to train using all the data. Under no circumstance should you report accuracy numbers of the classifier using training data without explicitly labeling the results as coming from training data; that is, the training data must not be mixed with the test data.

Prob. 12.7 — What are the two most popular approaches to cross-validation?

Answer (Prob. 12.7) — The two most popular approaches to cross-validation are k-fold cross validation, where $k = 5$ or $k = 10$, and leave-one-out cross validation, where $k = n$. (n is the number of data samples.)

Prob. 12.8 — Briefly explain the bias-variance tradeoff.

Answer (Prob. 12.8) — Allowing a more complex model potentially yields lower training error but oftentimes poorer generalization performance. Stated another way, more complex models yield less bias (that is, less expected discrepancy between the sample error and the true error) but higher variance (that is, higher expected variation in the error of the model when applied to new datasets).

Prob. 12.9 — Derive Equation (12.8).

Answer (Prob. 12.9) — Let $\tilde{f}$ be the true classifier. Assume that any data to which we have access is corrupted by zero-mean noise with variance σ^2, so that $\omega = \tilde{f}(\mathbf{x}) + \xi$, where ξ is a random variable such that $E[\xi] = 0$ and $E[\xi^2] = \sigma^2$. Our goal is to find a classifier f that approximates the true classifier $\tilde{f}$. Since all the data are corrupted by noise, we can do no better than find a classifier that deviates from the true classifier by the irreducible error σ^2.
For brevity, we shall drop the dependence upon $\mathbf{x}$, so that f represents $f(\mathbf{x})$ and $\tilde{f}$ represents $\tilde{f}(\mathbf{x})$. Given an unknown input $\mathbf{x}$ and its corresponding ground truth output ω, the squared error is given by $(f(\mathbf{x}) - \omega)^2$, or more simply, $(f - \omega)^2$. Over all possible inputs, then, the mean squared error is

$$
\begin{aligned}
E[(f-\omega)^2] &= E[f^2 - 2f\omega + \omega^2] && \triangleright \text{ Expand} \\
&= E[f^2] - 2E[f\omega] + E[\omega^2] && \triangleright\, E[X+Y] = E[X] + E[Y] \\
&= E[f^2] - 2E[f(\tilde{f} + \xi)] + E[(\tilde{f} + \xi)^2] && \triangleright\, \omega = \tilde{f} + \xi \\
&= E[f^2] - 2E[f\tilde{f} + f\xi] + E[\tilde{f}^2 + 2\tilde{f}\xi + \xi^2] && \triangleright \text{ Expand} \\
&= E[f^2] - 2E[f\tilde{f}] - 2E[f\xi] + E[\tilde{f}^2] + 2E[\tilde{f}\xi] + E[\xi^2] && \triangleright\, E[X+Y] = E[X] + E[Y] \\
&= E[f^2] - 2\tilde{f}E[f] - 2E[f\xi] + \tilde{f}^2 + 2\tilde{f}E[\xi] + E[\xi^2] && \triangleright\, E[\tilde{f}] = \tilde{f} \\
&= E[f^2] - 2\tilde{f}E[f] + \tilde{f}^2 + \sigma^2 && \triangleright\, E[\xi] = 0 \text{ and } E[\xi^2] = \sigma^2 \\
&= \mathit{Var}(f) + (E[f])^2 - 2\tilde{f}E[f] + \tilde{f}^2 + \sigma^2 && \triangleright\, \mathit{Var}(f) = E[f^2] - (E[f])^2 \\
&= \mathit{Var}(f) + (E[f] - \tilde{f})^2 + \sigma^2 && \triangleright \text{ Combine terms} \\
&= \mathit{Var}(f) + (E[f - \tilde{f}])^2 + \sigma^2 && \triangleright\, E[\tilde{f}] = \tilde{f} \\
&= \underbrace{(E[f - \tilde{f}])^2}_{\text{bias}} + \underbrace{\mathit{Var}(f)}_{\text{variance}} + \underbrace{\sigma^2}_{\text{irreducible error}} && \triangleright \text{ Rearrange}
\end{aligned}
$$

Note that in the 6^{th} and 10^{th} lines, $E[\tilde{f}] = \tilde{f}$ arises because $\tilde{f}$ is deterministic. The justification in the 8^{th} line is well-known and easy to prove from the definition of variance, letting $\mu_f \equiv E[f]$:

$$
\begin{aligned}
\mathit{Var}(f) &\equiv E[(f - \mu_f)^2] \\
&= E[f^2 - 2f\mu_f + \mu_f^2] \\
&= E[f^2] - 2E[f\mu_f] + E[\mu_f^2] \\
&= E[f^2] - 2\mu_f E[f] + \mu_f^2 \\
&= E[f^2] - 2\mu_f(\mu_f) + \mu_f^2 \\
&= E[f^2] - 2\mu_f^2 + \mu_f^2 \\
&= E[f^2] - \mu_f^2 \\
&= E[f^2] - (E[f])^2.
\end{aligned}
$$

Prob. 12.10 — Suppose you have a reasonably-sized dataset, and you compute both the Akaike and Bayesian information criteria (AIC and BIC). Which one do you expect to be greater?

Answer (Prob. 12.10) — For any reasonably-sized dataset, BIC > AIC.

Prob. 12.11 — Briefly explain the concept of structural risk minimization.

Answer (Prob. 12.11) — Structural risk minimization is a way to balance the complexity of a model against its ability to yield low training error.

Prob. 12.12 — What is the VC dimension of a classifier with a parabola-shaped decision boundary (at any orientation) in 2D?

Answer (Prob. 12.12) — The VC dimension of a classifier with a parabola-shaped decision boundary (at any orientation) in 2D is 4. To see this, note that any two points in the plane can be separated by some parabola. Therefore, the parabola shatters all pairs of points. Similarly, any three points in the plane can be separated by some parabola. Therefore, the parabola shatters all triples of points. Similarly, any four points in the plane can be separated by some parabola. Therefore, the parabola shatters all quadruples of points. However, it is easy to devise a set of five points that cannot be separated by any parabola. Therefore, since the VC dimension is the largest number of points that are shattered by the classifier space, the VC dimension is 4.

Prob. 12.13 — Explain how the curse of dimensionality, the peaking phenomenon, and feature selection are all related.

Answer (Prob. 12.13) — Because of the curse of dimensionality, it is not possible to gather sufficient training data to fill the space if more than a trivial number of dimensions are involved. As a result, adding dimensions to the feature vector may reduce performance (the peaking phenomenon). Feature selection is one way to overcome the peaking phenomenon, by carefully selecting features in the feature vector according to their performance.

Prob. 12.14 — Suppose we have a binary classification problem with the following training data, using 2D features:

category	set of 2D features (x_1, x_2)
ω_1	$\{(-12, 3)\}$
ω_2	$\{(10, 5), (-5, 5), (9, 6), (13, 0)\}$

Assume that the discrimination functions returned by the training procedure are the following:

$$\begin{aligned} g_1(x) &= 0.4x_1 + 0.5x_2 - 10 \\ g_2(x) &= 0.5x_1 + 0.3x_2 - 9 \end{aligned}$$

Calculate the accuracy of classification.

Answer (Prob. 12.14) — The classifier yields the correct result for 4 of the 5 data points. Therefore, the accuracy is $4/5 = 80\%$.

Prob. 12.15 — Suppose 145 test objects arrive in the detection zone of a conveyor belt, of which 95 are bananas. When a classifier is applied to these objects, 16 of the bananas are mislabeled, and 3 of the other objects are mislabeled as bananas.

(a) Compute TPR, TNR, FPR, and FNR, and build the confusion matrix.

(b) Calculate the sensitivity, specificity, and accuracy.

(c) Calculate the F-measure and the Jaccard coefficient.

Answer (Prob. 12.15) — The answers are as follows:

(a) The confusion matrix is

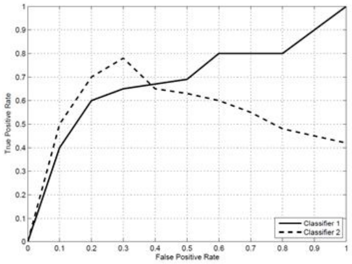
	test positive (banana detected)	test negative (not detected)
actual positive (banana)	79 (TP)	16 (FN)
actual negative	3 (FP)	47 (TN)

From this table, $TPR = 79/95 = 0.83$, $TNR = 47/50 = 0.94$, $FPR = 3/50 = 0.06$, $FNR = 16/95 = 0.17$.

(b) Sensitivity$= TPR = 0.83$, specificity$= TNR = 0.94$, accuracy$= (79 + 47)/145 = 0.87$.

(c) F-measure$= 79/(2 \cdot 79 + 3 + 16) = 0.45$, Jaccard$= 79/(79 + 3 + 16) = 0.81$.

Prob. 12.16 — Given the Receiver Operating Characteristic (ROC) curves of two classifiers below, find the better classifier using (a) the equal error rate (EER), and (b) the area under the curve (AUC).



Answer (Prob. 12.16) — (a) The EER of Classifier 1 is approximately 0.65, whereas the EER of Classifier 2 is approximately 0.75. Thus, according to EER, Classifier 2 is better. (b) From visual inspection, it is obvious that the AUC of Classifier 1 is greater than that of Classifier 2. Thus, according to AUC, Classifier 1 is better. However, note that either Classifier 2 was programmed incorrectly, or its ROC curve was measured incorrectly, its ROC curve does not intersect the point $(FPR, TPR) = (1, 1)$: even with 100% false positives, all of the true samples are not detected correctly.

Prob. 12.17 — When is a precision-recall (PR) curve preferred to a receiver operating characteristic (ROC) curve?

Answer (Prob. 12.17) — A PR curve is preferred to an ROC curve when the dataset is skewed, that is, when the number of items in the two categories are imbalanced. An example would be face detection, because images generally contain millions of pixels but only a few faces (at most).

Prob. 12.18 — Suppose that the population within a certain region is 51% male, and that 42% of the males wear eyeglasses, while 52% of the females wear eyeglasses. One person is randomly selected for a survey, and after the person is selected it is later learned that the person wears eyeglasses. Use Bayes' rule to calculate the probability that the selected person is male.

Answer (Prob. 12.18) — Let m = male, f = female, and e = eyeglasses. Then $p(m) = 0.51$, $p(e|m) = 0.42$, $p(e|f) = 0.52$. From Bayes' rule,

$$p(m|e) = \frac{p(e|m)p(m)}{p(e)} = \frac{p(e|m)p(m)}{p(e|m)p(m) + p(e|f)p(f)} = \frac{(0.42)(0.51)}{(0.42)(0.51) + (0.52)(0.49)} = 0.46.$$

The probability is therefore 46%.

Prob. 12.19 — Given the means and covariance matrices of the Gaussian densities of three categories ω_1, ω_2, and ω_3 as

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \qquad \boldsymbol{\mu}_3 = \begin{bmatrix} 6 \\ 9 \end{bmatrix} \qquad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix},$$

and the prior probabilities as $p(\omega_1) = 0.5$, $p(\omega_2) = p(\omega_3) = 0.25$, find the discriminant functions for each category, then calculate the decision boundary between the first two categories.

Answer (Prob. 12.19) — Since the Gaussian densities share the same covariance matrix, and since all the variances are equal, the discriminant functions are given by Equations (12.60) and (12.61). That is, the discriminant functions are $g_i(\mathbf{x}) \equiv \mathbf{w}_i^{\mathsf{T}}\mathbf{x} + \beta_i$, where

$$\mathbf{w}_1 = \frac{1}{10}\begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.1 \\ 0.1 \end{bmatrix} \tag{12.175}$$

$$\beta_1 = -\frac{1+1}{20} + \log 0.5 = -0.1 - 0.6931 = -0.7931 \tag{12.176}$$

$$\mathbf{w}_2 = \frac{1}{10}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0.3 \\ 0.4 \end{bmatrix} \tag{12.177}$$

$$\beta_2 = -\frac{9+16}{20} + \log 0.25 = -1.25 - 1.3863 = -2.6363 \tag{12.178}$$

$$\mathbf{w}_2 = \frac{1}{10}\begin{bmatrix} 6 \\ 9 \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.9 \end{bmatrix} \tag{12.179}$$

$$\beta_2 = -\frac{36+81}{20} + \log 0.25 = -5.85 - 1.3863 = -7.2363 \tag{12.180}$$

The decision boundary between the first two categories is given by $\mathbf{w}^{\mathsf{T}}(\mathbf{x}-\mathbf{c}) = 0$, where from Equations (12.62) and (12.63),

$$\begin{aligned}
\mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}\begin{bmatrix} -1-3 \\ 1-4 \end{bmatrix} = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}\begin{bmatrix} -4 \\ -3 \end{bmatrix} = \begin{bmatrix} -0.4 \\ -0.3 \end{bmatrix} \\
\mathbf{c} &= \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)} \log\frac{p(\omega_1)}{p(\omega_2)} \\
&= \frac{1}{2}(\begin{bmatrix} 2 \\ 5 \end{bmatrix}) + \frac{\begin{bmatrix} -4 \\ -3 \end{bmatrix}}{\begin{bmatrix} -4 & -3 \end{bmatrix}\begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}\begin{bmatrix} -4 \\ -3 \end{bmatrix}} \\
&= \begin{bmatrix} 1 \\ 2.5 \end{bmatrix} + \frac{\begin{bmatrix} -4 \\ -3 \end{bmatrix}}{2.5} \\
&= \begin{bmatrix} 1 \\ 2.5 \end{bmatrix} + \begin{bmatrix} -1.6 \\ -1.2 \end{bmatrix} \\
&= \begin{bmatrix} -0.6 \\ 1.3 \end{bmatrix}
\end{aligned}$$

Prob. 12.20 — Is Bayes' rule limited to Gaussian distributions? Why or why not?

Answer (Prob. 12.20) — Bayes' rule is applicable to any probability distribution. It is not limited to Gaussian distributions. Nothing in the derivation of Bayes' rule requires Gaussian distributions.

Prob. 12.21 — When is the maximum a posteriori (MAP) estimate the same as the maximum likelihood (ML) estimate?

Answer (Prob. 12.21) — The maximum a posteriori (MAP) estimate is the same as the maximum likelihood (ML) estimate when the prior is uniform. That is, $p(a|b)p(b) = p(a|b)$ when $p(b) = 1$.

Prob. 12.22 — True or false: For any value in the domain, the sum of all the class-conditional densities evaluated at that value equals 1.

Answer (Prob. 12.22) — False. If "class-conditional" is replaced by "posterior", then the statement is true.

Prob. 12.23 — Name one parametric and one nonparametric approach to representing a probability distribution.

Answer (Prob. 12.23) — One parametric approach involves Gaussian densities. Nonparametric approaches include histograms or Parzen windows over the raw data.

Prob. 12.24 — When are two Gaussian densities separated by a hyperplane?

Answer (Prob. 12.24) — Two Gaussian densities are separated by a hyperplane when they share the same covariance matrix.

Prob. 12.25 — Explain the difference between a generative method and a discriminative method for classification.

Answer (Prob. 12.25) — A generative method provides an explicit model for the probability density of each category, whereas a discriminative method only models the boundary between two categories.

Prob. 12.26 — Suppose you are given the following set of 1D training data along the x axis: $\{7, 12, 13, 15, 16\}$. Find the Parzen probability density function (PDF) estimate at $x = 15$, using the Gaussian function with variance 1 as the window function.

Answer (Prob. 12.26) — Using Gaussian functions, the Parzen PDF is

$$p(x) = \frac{1}{5}\sum_{i=1}^{5} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}},$$

where $\mu_1 = 7$, $\mu_2 = 12$, and so forth. With $\sigma^2 = 1$, this is

$$p(x) = \frac{1}{5\sqrt{2\pi}} \sum_{i=1}^{5} e^{-\frac{(x-\mu_i)^2}{2}}.$$

Evaluated at $x = 15$, the result is

$$\begin{aligned} p(15) &= \frac{1}{5\sqrt{2\pi}}(e^{-\frac{(15-7)^2}{2}} + e^{-\frac{(15-12)^2}{2}} + e^{-\frac{(15-13)^2}{2}} + e^{-\frac{(15-15)^2}{2}} + e^{-\frac{(15-16)^2}{2}}) \\ &= 0.1596(0 + 0.0111 + 0.1353 + 1 + 0.6065) \\ &= 0.2798 \end{aligned}$$

Prob. 12.27 — Explain how Parzen windows are related to locally weighted averaging (LWA).

Answer (Prob. 12.27) — LWA is the Parzen window PDF normalized by the evidence.

Prob. 12.28 — Following Section 12.3.1, collect a dataset of images and label most of the pixels as either red or not red using a paint program. (It is best to leave pixels that are ambiguous as unlabeled, and to not use them for training.) Write a program to construct positive and negative histograms of the class-conditional densities, then to find all the red pixels in a query image using this model.

Answer (Prob. 12.28) — Answers may vary.

Prob. 12.29 — List some strengths and weaknesses of nearest-neighbor classification.

Answer (Prob. 12.29) — Nearest-neighbor classification requires zero training time, and adding new data to the dataset is trivial. However, at run time, nearest-neighbor classification requires computation and memory related to the size of the dataset.

Prob. 12.30 — Construct a Naive Bayes classifier, using a Gaussian assumption, to classify whether a piece of fruit is an apple or plumb based on the measured features, including weight and perimeter. The training data set is provided below.

category	weight (g)	perimeter (cm)
apple	450	10.5
apple	332	9.6
apple	289	8.2
apple	265	8.3
apple	306	8.5
plumb	320	9.0
plumb	235	8.1
plumb	226	8.1
plumb	308	8.7
plumb	266	8.3

Apply the classifier to a test sample whose weight is 220 g and perimeter is 8.2 cm.

Answer (Prob. 12.30) — Let w be the weight, ϕ be the perimeter, a refer to apple, and b refer to plumb. Then

$$
\begin{aligned}
p(w|a) &= \frac{1}{\sqrt{2\pi\sigma^2_{w|a}}} e^{-\frac{-(w-\mu_{w|a})^2}{2\sigma^2_{w|a}}} \\
p(\phi|a) &= \frac{1}{\sqrt{2\pi\sigma^2_{w|a}}} e^{-\frac{-(\phi-\mu_{w|a})^2}{2\sigma^2_{w|a}}} \\
p(w|b) &= \frac{1}{\sqrt{2\pi\sigma^2_{w|a}}} e^{-\frac{-(w-\mu_{w|a})^2}{2\sigma^2_{w|a}}} \\
p(\phi|b) &= \frac{1}{\sqrt{2\pi\sigma^2_{w|a}}} e^{-\frac{-(\phi-\mu_{w|a})^2}{2\sigma^2_{w|a}}}
\end{aligned}
$$

where

$$\begin{aligned}
\mu_{w|a} &= \frac{1}{5}(450 + 332 + 2389 + 265 + 306) = 328.4 \\
\mu_{\phi|a} &= \frac{1}{5}(10.5 + 9.6 + 8.2 + 8.3 + 8.5) = 9.0 \\
\mu_{w|b} &= \frac{1}{5}(320 + 235 + 226 + 308 + 266) = 271.0 \\
\mu_{\phi|b} &= \frac{1}{5}(9.0 + 8.1 + 8.1 + 8.7 + 8.3) = 8.4 \\
\sigma^2_{w|a} &= 5218.3 \\
\sigma^2_{\phi|a} &= 0.997 \\
\sigma^2_{w|b} &= 1779.0 \\
\sigma^2_{\phi|b} &= 0.1580
\end{aligned}$$

Prob. 12.31 — Given the following set of 10 samples in a 3D space, follow the steps of PCA to calculate the eigenvalues of the orthogonal transformed data. Suppose the threshold of the fraction of the captured variance is set as 95%, show whether it is possible to reduce the dimensionality of the data.

$$\{(7,4,5),(6,5,4),(8,4,1),(2,6,9),(3,6,6),(5,7,3),(3,5,9),(2,8,6),(1,7,5),(8,5,2)\}$$

Answer (Prob. 12.31) — The covariance matrix is

$$\mathbf{C} = \begin{bmatrix} 6.9444 & -2.6111 & -5.2222 \\ -2.6111 & 1.7889 & 0.8889 \\ -5.2222 & 0.8889 & 7.1111 \end{bmatrix}$$

The eigenvalues of the covariance matrix are 12.8091, 2.7571, and 0.2782. The first eigenvalue alone captures $12.8/(12.8 + 2.8 + 0.3) = 80.5\%$ of the variance. The first two eigenvalues capture $(12.8 + 2.8)/(12.8 + 2.8 + 0.3) = 98.1\%$ of the variance. Therefore, the third dimension can be discarded while still retaining at least 95% of the variance.

Prob. 12.32 — What is the scree test?

Answer (Prob. 12.32) — The scree test determines the number of dimensions to retain by finding the knee in the eigenvalue plot.

Prob. 12.33 — Explain how active shape models (ASMs) and active appearance models (AAMs) are related.

Answer (Prob. 12.33) — An ASM models the shape of an object using PCA over the points defining the object. An AAM models both the shape and appearance of an object using PCA over the points and gray levels.

Prob. 12.34 — Calculate the linear discriminant function using Fisher's linear discriminant (FLD) for the following 2D data sets:

$$\mathcal{D}_1 = \{(5,3),(2,6),(3,5),(3,6),(4,7)\}$$
$$\mathcal{D}_2 = \{(7,9),(6,7),(9,5),(8,8),(10,8)\}$$

For simplicity, set the bias to the point halfway between the projected means.

Answer (Prob. 12.34) — The means are

$$\begin{aligned}\boldsymbol{\mu}_1 &= \begin{bmatrix}3.4\\5.4\end{bmatrix}\\ \boldsymbol{\mu}_2 &= \begin{bmatrix}8.0\\7.4\end{bmatrix}\end{aligned}$$

The scatter matrices are

$$\begin{aligned}\mathbf{S}_1 &= \begin{bmatrix}63 & 88\\88 & 155\end{bmatrix}\\ \mathbf{S}_2 &= \begin{bmatrix}330 & 294\\294 & 283\end{bmatrix}\end{aligned}$$

The within-class scatter matrix is

$$\mathbf{S}_w = \begin{bmatrix}393 & 382\\382 & 438\end{bmatrix}$$

Therefore, the normal vector (that is, the vector normal to the decision boundary) and bias (the point halfway between the projected means) are given by

$$\begin{aligned}\mathbf{w} &= \begin{bmatrix}-0.79\\0.61\end{bmatrix}\\ b &= 0.5771\end{aligned}$$

Prob. 12.35 — Implement (a) the perceptron learning algorithm with accelerated learning rate in Algorithm 12.2 and the pocket perceptron learning algorithm in Algorithm 12.3. Apply each algorithm to the datasets of the previous problem.

Answer (Prob. 12.35) — Answers may vary.

Prob. 12.36 — Use Lagrange multipliers to find the maximum and minimum values of the function $f(x, y) = 8x^2 + 200y^4$ with the constraint $x^2 + 2y^2 = 8$.

Answer (Prob. 12.36) — The function is shaped like an upward bowl centered at the origin. The constraint is an ellipse, also centered at the origin, that is longer in the x direction. As a result, we expect the maximum of the function subject to the constraint to occur at four points symmetric about the origin.
To use Lagrange multipliers, note that the constraint can be written as $g(x, y) = x^2 + 2y^2 - 8 = 0$. The gradients are

$$\begin{aligned}\nabla f(x,y) &= \begin{bmatrix}\partial f/\partial x & \partial f/\partial y\end{bmatrix}^\mathsf{T} = \begin{bmatrix}16x & 800y^3\end{bmatrix}^\mathsf{T}\\ \nabla g(x,y) &= \begin{bmatrix}\partial g/\partial x & \partial g/\partial y\end{bmatrix}^\mathsf{T} = \begin{bmatrix}2x & 4y\end{bmatrix}^\mathsf{T}\end{aligned}$$

Setting

$$\nabla f(x,y) = \lambda \nabla g(x,y)$$

we obtain

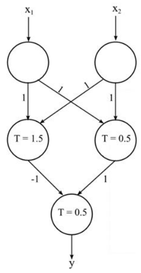
$$\begin{aligned}16x &= \lambda 2x\\ 800y^3 &= \lambda 4y\end{aligned}$$

Combining these equations yields $\lambda = 8$ and therefore $800y^2 = 32$, or $y = \pm 0.2$. Plugging these values into the constraint yields $x^2 + 2(0.04) = 8$, or $x \approx \pm 2.81$. The function evaluated at these points is $f(\pm 2.81, \pm 0.2) = 8(2.81)^2 + 200(0.2)^4 \approx 63.68$.

Prob. 12.37 — (a) What are the key ideas that turn a maximum margin classifier into a support vector machine (SVM)? (b) Explain the kernel trick.

Answer (Prob. 12.37) — (a) The key ideas that turn a maximum margin classifier into an SVM are the primal/dual formulation, slack variables, and the kernel trick. (b) The kernel trick takes advantage of the fact that the raw data do not appear in either the training or runtime equations of the classifier. As a result, a kernel can be applied to data pairs before further processing. With the appropriate kernel, the dimensionality is increased, so that a dataset that is not linearly separable in the original space can be linearly separable in the higher-dimensional space.

Prob. 12.38 — What logic operation does the following neural network perform? The weights of each branch are marked near the arrows, and the thresholds are as shown. The neurons output true if the input values pass the threshold, otherwise false.



Answer (Prob. 12.38) — This network implements the XOR function.

Prob. 12.39 — Plot the response of a logistic function, and of a rectified linear unit (ReLU). List some of the advantages of the latter.

Answer (Prob. 12.39) — The logistic function is an S-shaped curve with odd symmetry around the point $(0, 0.5)$; it approaches the value of 0 toward the left (as $x \to -\infty$) and the value of 1 toward the right (as $x \to \infty$). It exhibits a nearly linear response near the origin and zero derivatives far from the origin. The ReLU is a horizontal straight line with a value of 0 for all negative inputs ($f(x) = 0$ if $x < 0$), and it is a slanted line with a slope of 1 for all positive inputs ($f(x) = x$ if $x \geq 0$). The ReLU truncates negative values and thus introduces sparsity; and its gradient is constant for all positive values which decreases training time.

Prob. 12.40 — Briefly explain the concepts of bagging and boosting, and their relationship to one another.

Answer (Prob. 12.40) — Bagging and boosting are two different approaches to ensemble learning, that is, learning with multiple classifiers. Bagging refers to training multiple classifiers independently by repeatedly resampling (with replacement) the training data, then aggregating the results of these classifiers to achieve more accurate results than is possible with a single classifier. Boosting, on the other hand, trains multiple weak classifiers sequentially by allowing the later classifiers to focus on the data samples that were difficult for the earlier classifiers to classify. Bagging is used to decrease variance and avoid overfitting, whereas boosting is used to decrease bias and overcome underfitting.

Prob. 12.41 — Draw a degenerate decision tree, labeling the leaf nodes with "yes", "no", or "maybe".

Answer (Prob. 12.41) — Answers may vary. The tree should show a sequence of nodes (circles) connected by lines labeled "maybe." The last circle should spawn two lines, one labeled "yes" and the other labeled "no." All other circles should have a separate line emanating from them labeled "no".

Prob. 12.42 — What is a model containing parts with spring-like connections called?

Answer (Prob. 12.42) — A pictorial structure represents an object as a collection of parts with spring-like connections between the parts. It is a type of deformable part model (DPM), where the energy terms measure the likelihood of the parts and their relationships.

Prob. 12.43 — Sketch the main components of a deep neural network. Explain, at a high level, how this approach works, including the principles of max pooling and dropout.

Answer (Prob. 12.43) — Answers may vary. The network should show alternating convolutional and pooling layers, followed by several fully connected layers. Max pooling implements receptive fields, accomplishing scale invariance as well as data reduction. Dropout is a technique to enforce sparsity in the network.

Prob. 12.44 — Implement your own face detector. Collect several hundred images of faces from the internet and label them by hand by clicking on the images to define rectangles around the faces. (Be sure that all rectangles have the same aspect ratio.) Then collect several hundred images that contain no faces. Write code that performs preprocessing to generate a feature vector for a rectangle, and perform a sliding window search that computes a score for each hypothesis by passing the feature vector to a machine learning algorithm of your choice (using either your own implementation or code you find online). Report the accuracy of the detector on both the training set and on a separate test data, plotting results on an ROC or PR curve. Then write what you learned from this exercise.

Answer (Prob. 12.44) — Answers may vary.