Problem 1.

Consider the following set of 6 data points:

$$\vec{x}^{(1)} = (2, 4, 4)^T \tag{1}$$
$$\vec{x}^{(2)} = (-1, 2, 1)^T \tag{2}$$
$$\vec{x}^{(3)} = (3, -3, 2)^T \tag{3}$$
$$\vec{x}^{(4)} = (0, -3, -3)^T \tag{4}$$
$$\tag{5}$$

In the below parts, your answers should be given as numbers. You may leave your answer as an unsimplified fraction or a decimal, if you prefer.

a) What is the (1,3) entry of the sample covariance matrix?

b) What is the (1,2) entry of the sample covariance matrix?

Problem 2.

Recall that the Gaussian Radial Basis Function (RBF) kernel is defined as:

$$\kappa(\vec{x}, \vec{x}') = \exp\left(-\gamma \|\vec{x} - \vec{x}'\|^2\right).$$

Suppose a kernel SVM is trained on a data set with the Gaussian RBF kernel, and a training accuracy of 80% is achieved. If you wish to increase the training accuracy, what should be done to the value of γ?

- ◯ Increase the value of γ.
- ◯ Decrease the value of γ.
- ◯ Keep the value of γ the same, as it will not have an effect on the training accuracy.

Problem 3.

You and your friend are both training regularized least squares models on the same training data set of 50 points. You'd like to check if your models are the same by comparing the weights you've learned, but you're using different machine learning libraries that do things slightly differently.

The documentation of the library that your friend is using says that the regularization parameter is λ, and that the library minimizes the regularized risk:

$$\frac{1}{n}\sum_{i=1}^{n}(\vec{w}\cdot\vec{x}^{(i)}-y_i)^2+\lambda\|\vec{w}\|^2.$$

On the other hand, your library uses a regularization parameter of C and minimizes the regularized risk:

$$C\sum_{i=1}^{n}(\vec{w}\cdot\vec{x}^{(i)}-y_i)^2+\|\vec{w}\|^2.$$

If your friend is using a regularization parameter of $\lambda = 3$, what value of C should you use to ensure that your models are equivalent (that is, to get the same weights)?

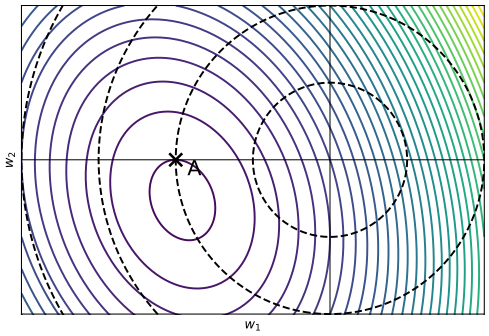
- ◯ $C = 3$
- ◯ $C = 1/3$
- ◯ $C = 150$
- ◯ $C = 1/150$
- ◯ It is not possible to guarantee that the models will be equivalent with any value of C.

Problem 4.

The 2-norm of a vector $\vec{w} \in \mathbb{R}^d$ is defined as:

$$\|\vec{w}\|_2 = \sqrt{\sum_{i=1}^{d} w_i^2}.$$

Let $R(\vec{w})$ be the *un*regularized empirical risk with respect to a data set. The solid curves below are the contours of $R(\vec{w})$. The dashed lines show where $\|\vec{w}\|_2$ is equal to 1, 2, 3, and so on.



Let $\tilde{R}(\vec{w}) = R(\vec{w}) + \lambda\|\vec{w}\|_2^2$ be the regularized empirical risk, with $\lambda > 0$.

True or False: there is a choice of λ so that the point marked A is a minimizer of the regularized empirical risk.

◯ True

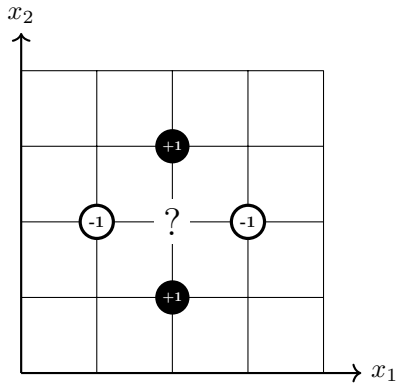
◯ False

Problem 5.

Consider the following data set of four points whose feature vectors are in $\mathbb{R}^2$ and whose labels are in $\{-1, 1\}$:

i	1	2	3	4
$\vec{x}^{(i)}$	(2,1)	(2,3)	(1,2)	(3,2)
y_i	1	1	-1	-1

For convenience, we've plotted the data below.



Suppose an unnamed kernel classifier $H(\vec{x}) = \sum_{i=1}^{n} \alpha_i \kappa(\vec{x}^{(i)}, \vec{x})$ has been trained on this data using a (spherical) Gaussian kernel and kernel width parameter $\gamma = 1$. Suppose the solution to the dual problem is found to be $\vec{\alpha} = (0, 4, -1, -2)^T$.

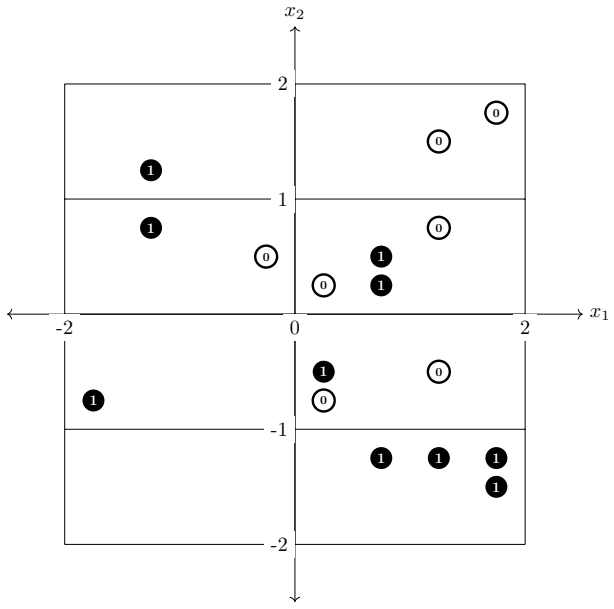
What class will the classifier predict for the point $(2, 2)$? For convenience, we've plotted this point on the graph above as a question mark.

◯ +1

◯ -1

Problem 6. (2 points)

For this problem, consider the binary classification training set shown below. The filled points have label 1 and the empty points have label 0. Suppose a histogram density estimator is used to estimate the class-conditional densities. The grid shown in the plot depicts the histogram bins; each bin has a width of 2 along the x_1 axis and height 1 along the x_2 axis.



a) What is the estimated density $p(x_1, x_2)$ at the point $(1.5, 1.5)$? You may leave your answer as a fraction or convert it to a decimal.

b) What is the estimated density $p_1(x_1, x_2 \,|\, Y = 1)$ at the point $(0.5, 0.5)$?

c) Using this histogram density estimator, what is the estimate of $\mathbb{P}(Y = 1 \,|\, x_1 = 0.5, x_2 = 0.5)$?

d) Suppose the Bayes classification rule is used with the estimated class-conditional densities in place of the true densities to predict the label of the point $(1, -.5)$. What is the predicted label?

◯ 1
◯ 0

Problem 7.

Consider the below data set collecting information on a set of 10 customers at a gym.

Personal Trainer?	Workout Type	Payment Method	Retained?
Yes	Yoga	Monthly	No
No	Cardio	Monthly	No
Yes	Cardio	Yearly	No
Yes	Strength	Monthly	No
No	Strength	Yearly	No
No	Cardio	Monthly	No
No	Yoga	Yearly	No
No	Yoga	Monthly	No
No	Strength	Yearly	Yes
Yes	Cardio	Yearly	Yes

Suppose we wish to predict whether a new customer will be retained. What does a Naïve Bayes classifier predict if the customer:

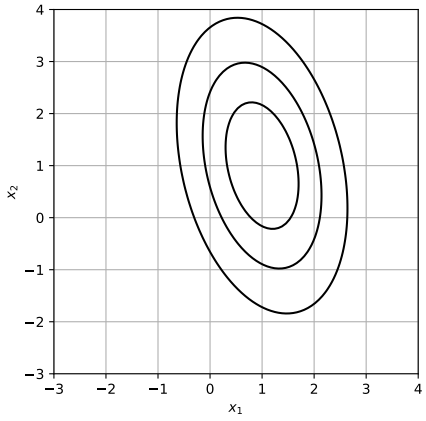
- Does not have a personal trainer,
- Does cardio workouts,
- Pays monthly?

◯ Retained

◯ Not Retained

Problem 8.

The plot below shows the contours of a multivariate Gaussian density:



Which of the following could be the Gaussian's covariance matrix?

- ◯ $\begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 3 \end{pmatrix}$
- ◯ $\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$
- ◯ $\begin{pmatrix} 3 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$
- ◯ $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Problem 9.

Why is it acceptable for AdaBoost to use weak learners, such as decision stumps, instead of strong learners?

- ◯ Because each weak learner is expected to classify every training point correctly.
- ◯ Because weak learners are simple and fast, and AdaBoost combines many of them into a stronger final classifier.
- ◯ Because weak learners ignore difficult points, which prevents overfitting.
- ◯ Because weak learners always generalize better than strong learners individually.

Problem 10. (3 points)

Suppose a discrete random variable X takes on values of either 0 or 1 and has the distribution:

$$\mathbb{P}(X = x) = \theta^x (1 - \theta)^{1-x}$$

where $\theta \in [0, 1]$ is a parameter.

Given a data set $x_1, \ldots, x_n$, what is the maximum likelihood estimate for the parameter θ? Show your work.

Problem 11.

A decision tree is being trained using Gini uncertainty. At the root node, two possible splits are being considered.

Split A produces one perfectly pure child node and one very mixed child node.

Split B produces two moderately pure child nodes.

Which statement is most accurate?

- ◯ Split A must be better because producing a pure child node is always the goal.
- ◯ Split B may be better because the tree evaluates the weighted uncertainty of both children, not just the purity of the best child.
- ◯ Split A and Split B are equally good according to Gini if they both reduce training error.
- ◯ The tree should choose whichever split uses the feature with the largest numerical values.

Problem 12.

True or False: Using a zero-mean Gaussian prior on the weights in linear regression leads to an L1 regularization penalty.

- ◯ True
- ◯ False

Problem 13.

A classifier achieves 5% test error on a test set of 40 samples.

True or False: The Bayes error for the distribution must be less than or equal to 5

- ◯ True
- ◯ False

Problem 14.

Suppose data are generated from a distribution with two well-separated clusters, but we fit a single Gaussian distribution using maximum likelihood.

As the number of samples grows, what should we expect?

- ◯ The fitted Gaussian must recover the true density exactly.
- ◯ The fitted Gaussian will converge to the best single-Gaussian approximation
- ◯ The fitted Gaussian will eventually split into two Gaussians.
- ◯ The fitted Gaussian will have zero variance.