CS 189 Introduction to Machine Learning
Summer 2018

FINAL EXAM

After the exam starts, **please write your student ID on every page**. There are **4** problems for a total of **143** points.
You may consult your two sheets of notes. Calculators, phones, computers, and other electronic devices are not permitted. There are **23** single-sided pages on the exam. **Notify a proctor immediately if a page is missing.** You may, without proof, use theorems and lemmas that were proven in the notes and/or in lecture, unless we explicitly ask for a derivation. **You have 80 minutes**.

We recommend you to leaf through every problem at the beginning of the exam to get a feel for the length. You do not have to work through the exam sequentially and should not spend excessive amounts of time on any particular question.

Exam Location: Lewis Hall, Room 100.

PRINT and SIGN Your Name: ____________________, ____________________, ____________________
(last) (first) (signature)

PRINT Your Student ID: ____

Person before you: ________________________________, ________________________________
(name) (SID)

Person behind you: ________________________________, ________________________________
(name) (SID)

Person to your left: ________________________________, ________________________________
(name) (SID)

Person to your right: ________________________________, ________________________________
(name) (SID)

Row (front is 1): ____________ Seat (leftmost is 1): ____________ (*Include empty seats/rows.*)

Do not turn this page until your instructor tells you to do so.

Extra page. If you want the work on this page to be graded, mention it on the problem’s main page.

1 Parameter Estimation (28 points)

1. (4 pts) Suppose we have n one-dimensional samples drawn i.i.d. from a continuous distribution with the following probability density function (PDF):

$$p(x;\mu) = \frac{1}{2}\exp\left(-|x-\mu|\right)$$

This distribution is special case of a family of continuous distributions known as Laplace distributions. In the questions below, we will derive the maximum likelihood estimate for this model.

Write down the log-likelihood function $\mathcal{L}(\mu)$ for this model in the form of a summation.

Solution:

$$\begin{aligned}\mathcal{L}(\mu) &= \log p(x_1,\ldots,x_n;\mu)\\ &= \log\prod_{i=1}^{n} p(x_i;\mu)\\ &= \sum_{i=1}^{n}\log\big(1/2\big) - |x_i-\mu|\\ &= n\log\big(1/2\big) - \sum_{i=1}^{n}|x_i-\mu|\end{aligned}$$

2. (4 pts) **Write down the derivative with respect to μ of the log-likelihood function from the previous part.** Recall that the maxima and minima of a function can occur at any place where the derivative is zero or undefined. You may use the following fact:

$$\frac{d\,|x|}{d\,x} = \begin{cases}\text{sign}(x) & \text{if } x \neq 0\\ \text{undefined} & \text{if } x = 0\end{cases} = \begin{cases}1 & \text{if } x > 0\\ -1 & \text{if } x < 0\\ \text{undefined} & \text{if } x = 0\end{cases}$$

Solution:

$$\begin{aligned}\frac{d\mathcal{L}(\mu)}{d\mu} &= \frac{d}{d\mu} n\log\big(1/2\big) - \sum_{i=1}^{n}|x_i-\mu|\\ &= -\sum_{i=1}^{n}\frac{d}{d\mu}|x_i-\mu|\\ &= \begin{cases}-\sum_{i=1}^{n}\text{sign}(\mu-x_i) & \text{if } \mu\neq x_i\ \forall i\\ \text{undefined} & \text{if } \exists i \text{ such that } \mu = x_i\end{cases}\end{aligned}$$

3. (5 pts) Consider the case when n is odd. **Argue that the maximum likelihood estimate of μ is the median of $x_1, \ldots, x_n$.** Hint: think about how the objective function (log-likelihood) changes as the estimate of μ moves towards the median.

Solution:

We can assume without loss of generality that $x_1 \leq \ldots \leq x_n$. So, the median is $x_{(n+1)/2}$.

When $\mu \neq x_i \forall i$, each term in the derivative, sign$(\mu - x_i)$, is 1 or -1. Because n is odd, the derivative, $-\sum_{i=1}^n$ sign$(\mu - x_i)$, cannot be zero. So the MLE cannot occur between any of the x_i's.

When $\mu = x_i$ for some i, the derivative is undefined, so the MLE could occur at any of the x_i's. Consider what happens to $\sum_{i=1}^n |x_i - \mu|$ if we move μ from x_j to x_{j+1}. Observe that $\sum_{i=1}^n |x_i - \mu| = \sum_{i=1}^j |x_i - \mu| + \sum_{i=j+1}^n |x_i - \mu|$. For each $i \leq j$, $|x_i - \mu|$ increases by $x_{j+1} - x_j$ and so the first term increases by $j(x_{j+1} - x_j)$. For each $i \geq j+1$, $|x_i - \mu|$ decreases by $x_{j+1} - x_j$ and so the second term decreases by $(n-j)(x_{j+1} - x_j)$. If $j < (n+1)/2$, $n - j > j$ and so the decrease in the second term is greater than the increase in the first term, and so the overall sum decreases. On the other hand, If $j \geq (n+1)/2$, $n - j < j$ and so the decrease in the second term is less than the increase in the first term, and so the overall sum increases.

Since we want to maximize $n \log(1/2) - \sum_{i=1}^n |x_i - \mu|$, we want to minimize $\sum_{i=1}^n |x_i - \mu|$. The above shows that as μ is moved towards $x_{(n+1)/2}$, $\sum_{i=1}^n |x_i - \mu|$ always decreases, and so the maximum occurs at $\mu = x_{(n+1)/2}$.

There is also another line of argument that conveys the right intuition, but is slightly incorrect:

If we choose $\mu = x_{(n+1)/2}$, then there are $(n-1)/2$ x_i's less than μ and $(n-1)/2$ x_i's greater than μ. So, $\sum_{i=1}^n$ sign$(\mu - x_i)$ is zero, if we define sign$(0) := 0$. If we choose $\mu = x_j$ where $j < (n+1)/2$, then there are more x_i's that are greater than μ than there are x_i's less than μ. And so $\sum_{i=1}^n$ sign$(\mu - x_i)$ is negative. If we choose $\mu = x_j$ where $j > (n+1)/2$, then $\sum_{i=1}^n$ sign$(\mu - x_i)$ is positive. So, the only case when the derivative, $\sum_{i=1}^n$ sign$(\mu - x_i)$, is zero is when $\mu = x_{(n+1)/2}$, implying that $\mu = x_{(n+1)/2}$ maximizes likelihood. This is technically incorrect because the derivative is not always $\sum_{i=1}^n$ sign$(\mu - x_i)$; it is in fact undefined if μ is equal to any one of the x_i's.

4. (5 pts) Now consider the case when n is even. **Argue that the maximum likelihood estimate of μ could be any number between the two central values of $x_1, \ldots, x_n$.** In particular, because the median in this case is defined as the average of the two central values, it is a maximum likelihood estimate of μ both when n is odd and when n is even.

Solution: Again assume without loss of generality that $x_1 \leq \ldots \leq x_n$.

When $\mu \neq x_i \forall i$, the derivative $-\sum_{i=1}^{n} \text{sign}(\mu - x_i)$ can only be zero if $x_{n/2} < \mu < x_{n/2+1}$. So the MLE could occur at any point between $x_{n/2}$ and $x_{n/2+1}$ exclusive.

When $\mu = x_i$ for some i, the derivative is undefined, so the MLE could also occur at any of the x_i's. By similar reasoning as in the previous part, we consider the effect of moving μ from x_j to x_{j+1} and break $\sum_{i=1}^{n} |x_i - \mu|$ into two terms: $\sum_{i=1}^{j} |x_i - \mu| + \sum_{i=j+1}^{n} |x_i - \mu|$. If $j < n/2$, $n - j > j$ and so the decrease in the second term is greater than the increase in the first term, and so the overall sum decreases. On the other hand, If $j \geq n/2 + 1$, $n - j < j$ and so the decrease in the second term is less than the increase in the first term, and so the overall sum increases. This shows that as μ is moved towards $x_{n/2}$ from below or towards $x_{n/2+1}$ from above, $\sum_{i=1}^{n} |x_i - \mu|$ always decreases. Since maximization of $n \log(1/2) - \sum_{i=1}^{n} |x_i - \mu|$ is equivalent to minimization of $\sum_{i=1}^{n} |x_i - \mu|$, log-likelihood could not be maximized at any of the x_i's other than $x_{n/2}$ and $x_{n/2+1}$.

Because $-\sum_{i=1}^{n} |x_i - \mu|$ is continuous and its derivative w.r.t. μ is zero on the interval $(x_{n/2}, x_{n/2+1})$, the value of $-\sum_{i=1}^{n} |x_i - \mu|$ is constant on the interval $[x_{n/2}, x_{n/2+1}]$. Therefore, any value of $\mu \in [x_{n/2}, x_{n/2+1}]$ maximizes likelihood, and so the median is an MLE.

5. (2 pts) Consider the following ordered 1-dimensional dataset with an outlier. **Which is greater: the mean or the median? Briefly explain why.**

$$D = \{1, 2, 4, 5, 6, 32\}$$

Solution:

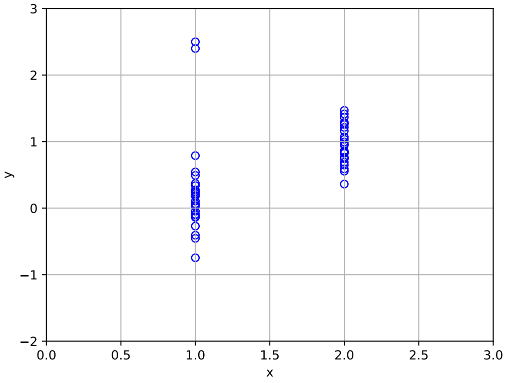
The mean is greater than the median. Notice that regardless of the value of the final point in the dataset (assuming it is greater than or equal to 6), the median will always be the average of the two central vales i.e. 4.5. In contrast a large outlier in the dataset will affect the mean. In this case the 32 drags the mean up.

6. (4 pts) Let $x, y \in \mathbb{R}$ be random variables denoting the (one-dimensional) input data and label respectively. Consider the following models:

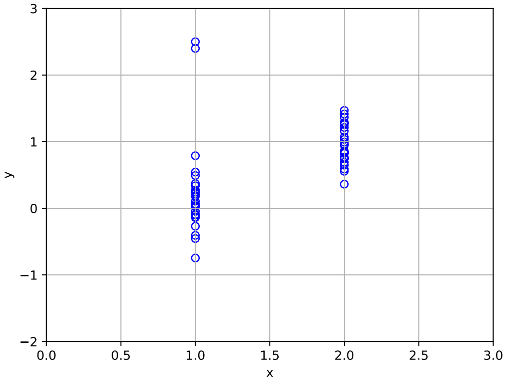
   (a) $p(y|x; w, b) = \frac{1}{2} \exp\big(-|y - (wx + b)|\big)$

   (b) $y|x \sim \text{Normal}(wx + b, 1)$, and so $p(y|x; w, b) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{(y-(wx+b))^2}{2}\Big)$

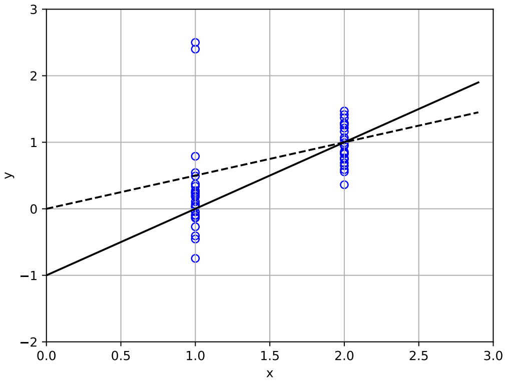
   Consider the dataset shown in the plot below. Suppose that for each model, we estimate parameters w and b using maximum likelihood estimation (MLE) on the dataset. Let w^* and b^* denote the maximum likelihood estimates of w and b respectively. **In the plot below, sketch the line $\hat{y}_{\text{test}} = w^* x_{\text{test}} + b^*$ for each model using a solid line for (a) and a dashed line for (b).**



If you make a mistake, cross out the plot above and draw in the plot below:



Solution:



7. (4 pts) **Briefly explain why the lines you drew in the previous part are positioned the way they are.**

Solution:

Since there are two parameters in our linear regression model and only two possible values of x's, the resulting lines in both (a) and (b) go through the maximum likelihood estimates of y at each value of x. For (b), recall that the MLE of the mean of a Gaussian is the sample mean, and so the line goes through the sample mean of the y-coordinates of points that have an x-coordinate of 1 and the sample mean of the y-coordinates of points that have an x-coordinate of 2. For $x = 1$, the value of the sample mean is affected by the outliers, which drag the mean upward and away from the center of the cluster. On the other hand, for (a), the line goes through the median of the y-coordinates of points that have an x-coordinate of 1 and the sample mean of the y-coordinates of points that have an x-coordinate of 2. Unlike sample means, medians are robust to outliers as seen in the previous part. Therefore, for $x = 1$, the value of the median is closer to the center of the cluster.

2 Neural Nets and Optimization (45 points)

1. Consider a fully-connected neural net with one input layer, one output layer and no hidden layers. Both the input and output layers contain only one neuron. The output layer has linear activations with no biases. The input and output dimensionalities are both 1. Let x denote the the input data, y denote the ground truth label, $\hat{y}$ denote the predicted label and w denote the weight from the input layer to the output layer. The model can therefore be expressed as $\hat{y} = wx$. It is trained using the ℓ_2 loss, $L = (y - \hat{y})^2$.

   (a) (3 pts) **Derive the expression for gradient with respect to the sole parameter w, i.e.: $\frac{\partial L}{\partial w}$. Then evaluate it when $w = 0$.** Show your work.

   Solution:

$$\begin{aligned} L &= (y - \hat{y})^2 \\ \frac{\partial L}{\partial w} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w} \\ &= -2(y - \hat{y}) \cdot \frac{\partial \hat{y}}{\partial w} \\ &= -2(y - \hat{y}) \cdot x \\ &= -2(y - wx) \cdot x \\ &= -2yx + 2wx^2 \end{aligned}$$

   If $w = 0$, then we have

$$L = -2yx + 2wx^2 = -2yx$$

   (b) (4 pts) Consider training the neural net on a dataset with a single example, where the input x and the label y are both 1. **What is the value of the loss function at a global minimum and what is a setting of the parameter w that attains it? Justify why it is a global minimum rather than just a local minimum or a critical point.**

   Solution:

   At the global optimum, the loss is 0. Since the loss is always non-negative because of the square, the minimum it can possibly achieve is zero. An example that realizes this minimum is setting

$$w = \frac{y}{x}$$

   As we can verify by doing:

$$\begin{aligned} \hat{y} = wx = \frac{y}{x}x = y \\ (y - \hat{y})^2 = (y - y)^2 = 0 \end{aligned}$$

   In our case, because $x = y = 1$, a setting of w that attains the global minimum is $w = y/x = 1/1 = 1$.

(c) (4 pts) **Show that the loss function is convex in the parameter w.**

Solution:

Recall that a twice-differential function is convex if its second derivative is always non-negative.

$$\begin{aligned}\frac{\partial^2 L}{\partial w^2} &= \frac{\partial L}{\partial w}(-2yx + 2wx^2) \\ &= 0 + 2x^2 \\ &\geq 0\end{aligned}$$

Since $2x^2 \geq 0$ for all x, the loss function is convex.

(d) (4 pts) **Show that the gradient of the loss function, $\frac{\partial L}{\partial w}$, is Lipschitz continuous.** Recall the definition of Lipschitz continuity: a function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous if there exists a constant $C > 0$ such that $\left|f(w_0) - f(w_1)\right| \leq C\left|w_0 - w_1\right|$ for all $w_0, w_1 \in \mathbb{R}$. Note that x and y are given and can therefore be treated as constants.

Solution: If we want to show that $\frac{\partial L}{\partial w}$ is Lipschitz continuous, we need to show that:

$$|\frac{\partial L}{\partial w} - \frac{\partial L}{\partial w'}| \leq C|w - w'|$$

Using part (a)

$$\begin{aligned}&|\frac{\partial L}{\partial w} - \frac{\partial L}{\partial w'}| \\ &= |-2yx + 2wx^2 - (-2yx + 2w'x^2)| \\ &= |2wx^2 - 2w'x^2| \\ &= 2x^2|w - w'| \\ &\leq C|w - w'|\end{aligned}$$

For some $C \geq 2x^2$. Note that this essentially follows from the fact that the gradient of a quadratic is linear.

(e) (4 pts) Starting from **the all-zeros initialization (when w is initialized to 0)**, consider running gradient descent with a constant learning rate that is small enough to ensure convergence. **What is the value of the loss function that gradient descent will converge to? Explain your reasoning.**

Solution: Gradient descent will reach the global minimum since the loss function is convex and its gradient is Lipschitz continuous. Parameters will slowly move away from being all 0 and towards the solution.

2. Consider a fully-connected neural net with one input layer, **one hidden layer** and one output layer. All layers contain only one neuron. Both the hidden and output layers have linear activations with no biases. The input and output dimensionalities are both 1. Let x denote the the input data, y denote the ground truth label, $\hat{y}$ denote the predicted label, v denote the weight from the input layer to the hidden layer and w denote the weight from the hidden layer to the output layer. The model can therefore be expressed as $\hat{y} = wvx$. It is trained using the ℓ_2 loss, $L = (y - \hat{y})^2$.

   (a) (6 pts) **Derive the expression for gradient with respect to the parameters (w, v), i.e.: $\nabla_{(w,v)}L := \left(\frac{\partial L}{\partial w}, \frac{\partial L}{\partial v}\right)$. Then evaluate it when $w = 0$ and $v = 0$.** Show your work. Hint: the answer may surprise you.

   Solution:

$$\begin{aligned} L &= (y - \hat{y})^2 \\ \frac{\partial L}{\partial w} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w} \\ &= -2(y - \hat{y}) \cdot \frac{\partial \hat{y}}{\partial w} \\ &= -2(y - \hat{y}) \cdot vx \\ &= -2(yvx - w(vx)^2) \\ &= -2v(yx - wvx^2) \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial v} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial v} \\ &= -2(y - \hat{y}) \cdot \frac{\partial \hat{y}}{\partial v} \\ &= -2(y - \hat{y}) \cdot wx \\ &= -2(ywx - v(wx)^2) \\ &= -2w(yx - vwx^2) \end{aligned}$$

   When $v = 0$, $\frac{\partial L}{\partial w} = 0$. When $w = 0$, $\frac{\partial L}{\partial v} = 0$. Hence, $\nabla L(0, 0) = (0, 0) = \mathbf{0}$.

   (b) (5 pts) Consider training the neural net on the same dataset as before (there is a single example, where the input and label are both 1). **What is the value of the loss function at a global minimum and what is a setting of the parameters (w, v) that attains it? Justify why it is a global minimum rather than just a local minimum or a critical point.**

   Solution:

   The loss at the global optimum is 0. Since the loss is squared, the minimum it can possibly achieve is zero. An example that realizes this minimum is any setting where $wv = \frac{y}{x} = 1$. To verify this, we just do:

$$\begin{aligned} L = (y - \hat{y})^2 = (y - wvx)^2 \\ &= (y - \frac{y}{x}x)^2 \\ &= (y - y)^2 = 0 \end{aligned}$$

   For example, both $w = 3, v = \frac{1}{3}$ and $w = 1, v = 1$ will work.

(c) (5 pts) **How many different settings of the parameters (w, v) attain the global minimum? Explain your reasoning. If there are more than one, write down the set of all parameter settings that attain the global minimum.** Hint: parameters can take on continuous values.

Solution:

Infinite. Any solution which satisfies $wv = 1$ attains the global minimum.

(d) (5 pts) Consider a parameter setting $w = 0$ and $v = 0$. **Is this a global minimum of L, a global maximum of L, or neither? Explain your reasoning.**

Solution:

From before, we know that the global minimum is 0. Also, note that
$L = (y - \hat{y})^2 = (1 - wv)^2 \to \infty$ as $wv \to \infty$ and thus it is unbounded from above.
However, if $w = v = 0$, then $(y - wvx)^2 = (1 - 0)^2 = 1$ which is neither of the above cases.

(e) (5 pts) Starting from **the all-zeros initialization (when w and v are both initialized to** 0), consider running gradient descent with a constant learning rate that is small enough to ensure convergence. **What is the value of the loss function that gradient descent will converge to? Explain your reasoning. Does gradient descent find a global minimum in this case?**

Solution:

Since the gradient is zero, we will not move from our initialization and will converge to $(0, 0)$, which we showed above is not a global minimum. To see this, write out the GD step of:

$$w^1 = w^0 - \eta(\frac{\partial L}{\partial w}|_{w^0})$$

$$= 0 - \eta(0)$$

$$= 0 = w^0$$

Proceed by induction and by replacing w with v.

3 A New Unsupervised Learning Method (40 points)

1. (4 pts) Consider the following optimization problem. The $\mathbf{x}_i$'s are the training examples, which are given. Note that there are no labels, so this is an unsupervised learning method. The variables m, $\mathbf{w}$ and b are those that we optimize over.

$$\begin{aligned} \min_{m,\mathbf{w},b} \quad & m \\ \text{s.t.} \quad & \frac{|\mathbf{w}^\top \mathbf{x}_i - b|}{\|\mathbf{w}\|_2} \leq m \quad \forall i \\ & m \geq 0 \end{aligned}$$

How is this different from a hard-margin SVM? Recall that a hard-margin SVM solves the following problem, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ are the ith training example and ground truth label respectively:

$$\begin{aligned} \max_{m,\mathbf{w},b} \quad & m \\ \text{s.t.} \quad & \frac{y_i \left(\mathbf{w}^\top \mathbf{x}_i - b\right)}{\|\mathbf{w}\|_2} \geq m \quad \forall i \\ & m \geq 0 \end{aligned}$$

Solution:

There are two key differences:

(a) The direction of inequality is flipped in the first constraint, which means we want all the points inside or on the margin.

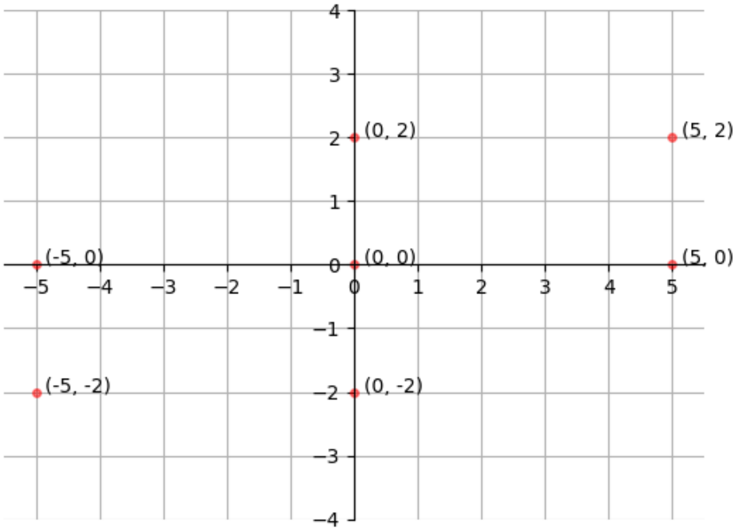
(b) The maximization is changed to minimization, which means we want to minimize the margin.

The third difference is already given in the question, which is that the new method does not use labels, whereas hard-margin SVM does.

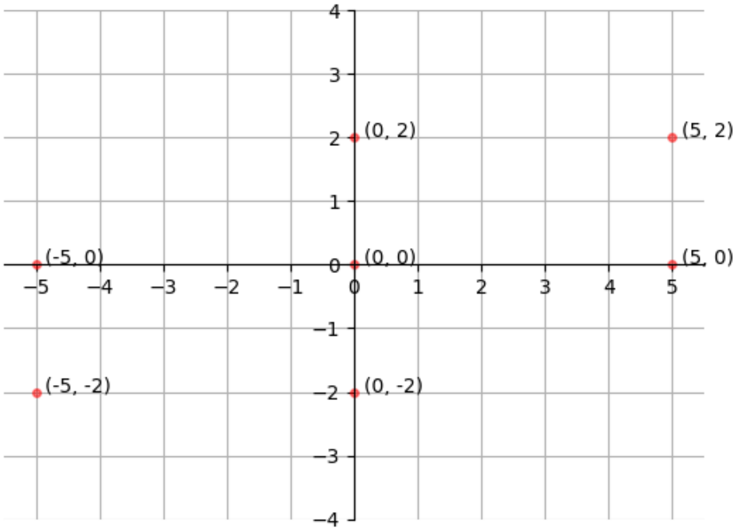
2. (4 pts) **What does $|\mathbf{w}^\top \mathbf{x}_i - b|/\|\mathbf{w}\|_2$ represent geometrically?**

Solution: $|\mathbf{w}^\top \mathbf{x}_i - b|/\|\mathbf{w}\|_2$ represents the perpendicular distance from $\mathbf{x}_i$ to the line $\mathbf{w}^{*\top} \mathbf{x}_{\text{test}} - b^* = 0$.

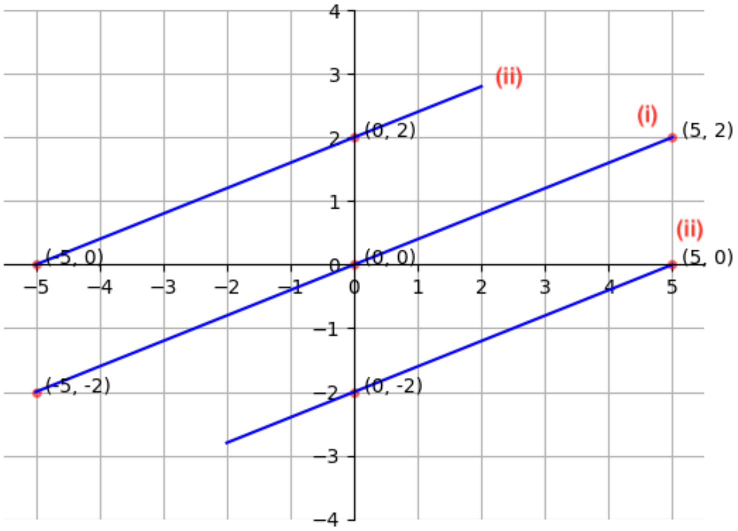
3. (6 pts) Draw the optimal solution for the 2D dataset shown in the plot below. More specifically, let $\mathbf{w}^*$, b^* and m^* denote the optimal solution. **Draw (i) the line such that $\mathbf{w}^{*\top}\mathbf{x}_{\text{test}} - b^* = 0$ and (ii) the two lines such that $\left(\mathbf{w}^{*\top}\mathbf{x}_{\text{test}} - b^*\right)/\|\mathbf{w}^*\|_2 = m^*$ and $\left(\mathbf{w}^{*\top}\mathbf{x}_{\text{test}} - b^*\right)/\|\mathbf{w}^*\|_2 = -m^*$. Indicate which line corresponds to which by labelling each line with either (i) or (ii).**



If you make a mistake, cross out the plot above and draw in the plot below:



Solution:



4. (4 pts) **Briefly explain why the lines you drew in the previous part are positioned the way they are.**

   Solution: The goal of the optimization problem in part 1 is to find a hyperplane with the minimum margin such that all the training points lie within the margin. Geometrically, (i) is the hyperplane and (ii) are the margin hyperplanes, which should be parallel and symmetric about (i), and the distance between (i) and (ii) should be minimized. A common mistake is to have the horizontal axis as (i); the margin in this case is wider than the margin pictured in the previous part, due to Pythagorean theorem.

5. (6 pts) Recall the optimization problem in part 1, which is reproduced below for convenience:

$$\begin{aligned} \min_{m,\mathbf{w},b} \quad & m \\ \text{s.t.} \quad & \frac{|\mathbf{w}^\top \mathbf{x}_i - b|}{\|\mathbf{w}\|_2} \le m \quad \forall i \\ & m \ge 0 \end{aligned}$$

   It turns out that this optimization problem can be equivalently expressed as

$$\begin{aligned} \max_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & |\mathbf{w}^\top \mathbf{x}_i - b| \le 1 \quad \forall i \end{aligned}$$

   where we remove m and just optimize over $\mathbf{w}$ and b. **Argue why the two optimization problems are equivalent.**

   Solution: We first show that constraining $\mathbf{w}$ to have a specific norm doesn't change the solution. Suppose that $(m^*, \mathbf{w}^*, b^*)$ is a solution to the optimization problem in part 1. Then, for any $k > 0$, $(m^*, k\mathbf{w}^*, kb^*)$ is also a solution. So, we can add a constraint $\|\mathbf{w}\|_2 = \frac{1}{m}$, or equivalently, $m = \frac{1}{\|\mathbf{w}\|_2}$:

$$\begin{aligned} \min_{m,\mathbf{w},b} \quad & m \\ \text{s.t.} \quad & \frac{|\mathbf{w}^\top x_i - b|}{\|\mathbf{w}\|_2} \le m \quad \forall i \\ & m \ge 0 \\ & m = \frac{1}{\|\mathbf{w}\|_2} \end{aligned}$$

   Substitute $m = \frac{1}{\|\mathbf{w}\|_2}$ and eliminate m, we have:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{\|\mathbf{w}\|_2} \\ \text{s.t.} \quad & |\mathbf{w}^\top x_i - b| \le 1 \quad \forall i \end{aligned}$$

   Equivalently,

$$\begin{aligned} \max_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & |\mathbf{w}^\top x_i - b| \le 1 \quad \forall i \end{aligned}$$

6. (6 pts) The dual to the optimization problem in the previous part is

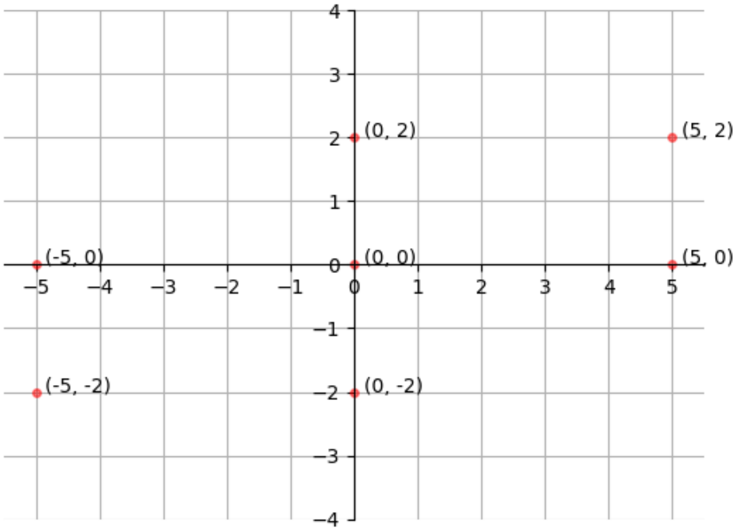
$$\max_{\boldsymbol{\alpha}\geq 0, \boldsymbol{\beta}\geq 0} \min_{\mathbf{w},b} -\frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_i \alpha_i(\mathbf{w}^\top \mathbf{x}_i - b - 1) + \sum_i -\beta_i(\mathbf{w}^\top \mathbf{x}_i - b + 1)$$

where we have

- Primal variables $\mathbf{w}$ and b
- Dual variables α_i corresponding to each constraint of the form $\mathbf{w}^\top x_i - b - 1 \leq 0$
- Dual variables β_i corresponding to each constraint of the form $-(\mathbf{w}^\top x_i - b + 1) \leq 0$

It turns out that strong duality holds in this case.

Draw and label the lines you drew in part 1 again in the plot below. Consider the set of $\mathbf{x}_i$'s such that either α_i or β_i is non-zero. **Circle all points that *could* belong to this set. Explain your reasoning.**

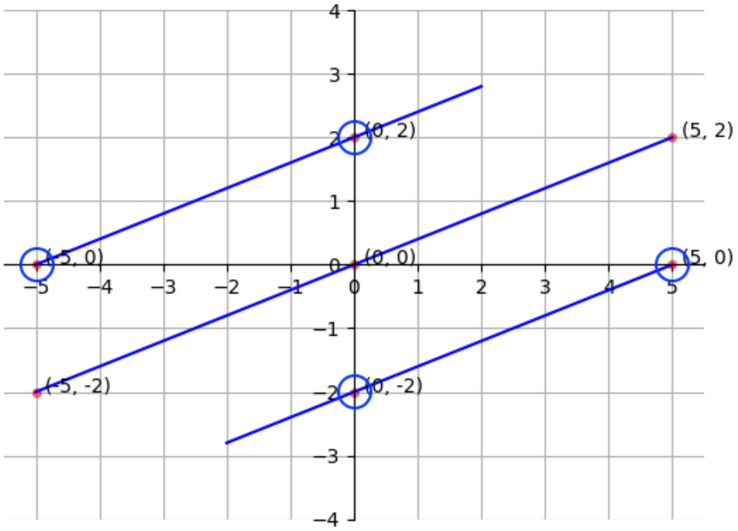


Solution:

Let $(\mathbf{w}^*, b^*)$ and $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ be the primal and dual solutions respectively, since strong duality holds, they satisfy the KKT conditions. Specifically, they satisfy the complementary slackness conditions, that is:

$$\alpha_i^*(\mathbf{w}^{*\top} x_i - b^* - 1) = 0 \quad \forall i$$
$$\text{and}$$
$$-\beta_i^*(\mathbf{w}^{*\top} x_i - b^* + 1) = 0 \quad \forall i$$

When $\alpha_i^* = 0$ or $\beta_i^* = 0$, it must be true that $\mathbf{w}^{*\top} x_i - b^* - 1 < 0$ or $\mathbf{w}^{*\top} x_i - b^* + 1 > 0$. Therefore, points corresponding to non-zero dual variables must lie exactly on the margin.


(0, 2)
(5, 2)
(-5, 0)
(0, 0)
(5, 0)
(-5, -2)
(0, -2)
−5
−4
−3
−2
−1
0
1
2
3
4
5
4
3
2
1
0
−1
−2
−3
−4

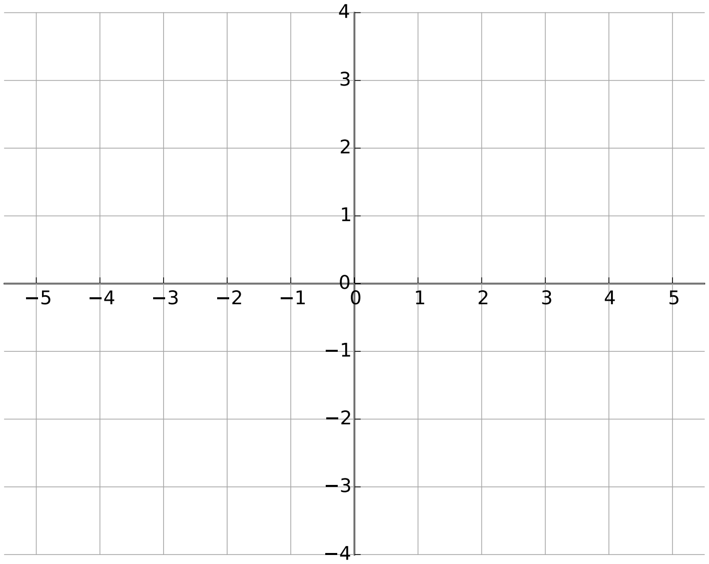
7. (6 pts) Let $\mathbf{w}^*$, b^* and m^* denote the optimal solution to the optimization problem in part 1. **In the plots below, draw two datasets with the following properties**:

   (a) A dataset with 10 examples such that $\mathbf{w}^*$ coincides with the first principal component

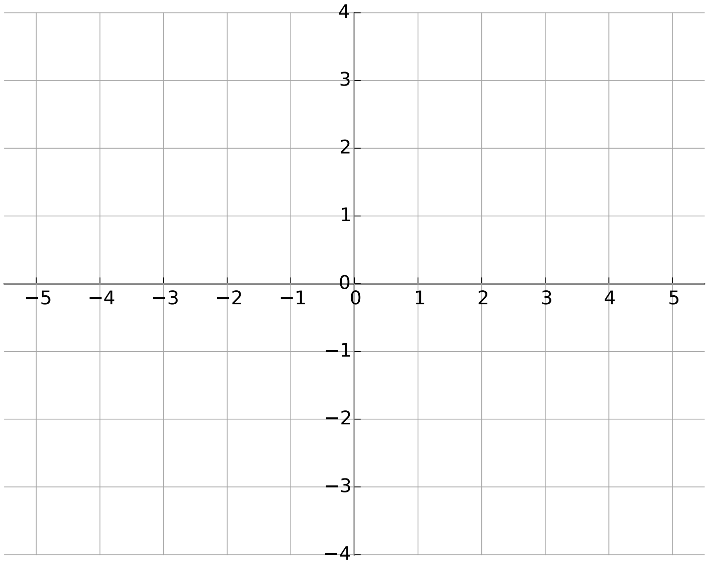
   (b) A dataset with the same 10 examples as in (a) and **one** extra example such that $\mathbf{w}^*$ differs from the first principal component

   In each case, also draw (i) the line such that $\mathbf{w}^{*\top}\mathbf{x}_{\text{test}} - b^* = 0$, (ii) the two lines such that $\left(\mathbf{w}^{*\top}\mathbf{x}_{\text{test}} - b^*\right)/\|\mathbf{w}^*\|_2 = m^*$ and $\left(\mathbf{w}^{*\top}\mathbf{x}_{\text{test}} - b^*\right)/\|\mathbf{w}^*\|_2 = -m^*$ and (iii) the first principal component, and label each with (i), (ii) or (iii). Use x's to indicate data points.

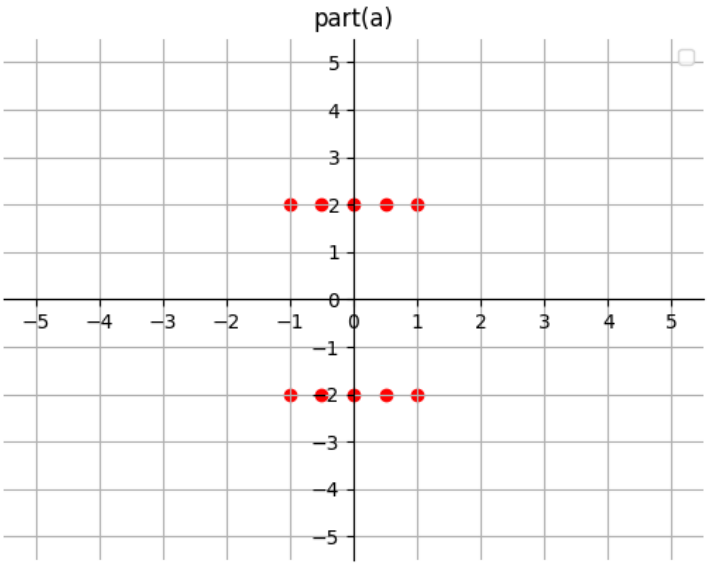
   (a) $\mathbf{w}^*$ same as the first principal component:

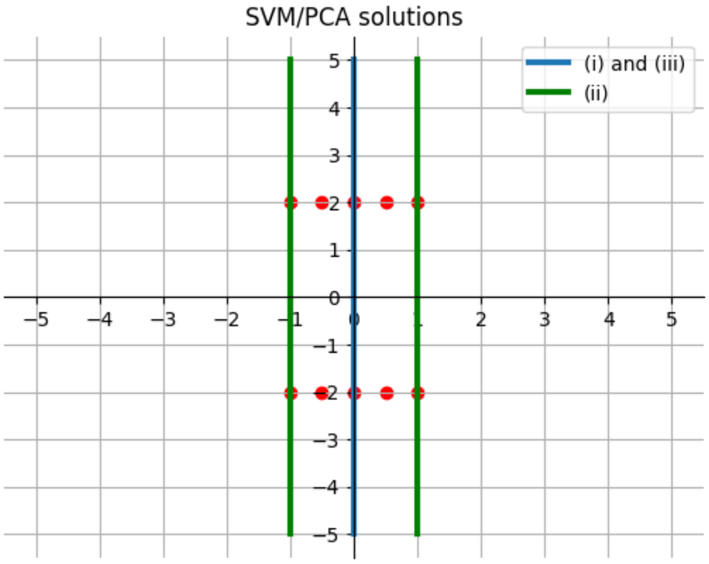


   (b) $\mathbf{w}^*$ different from the first principal component:

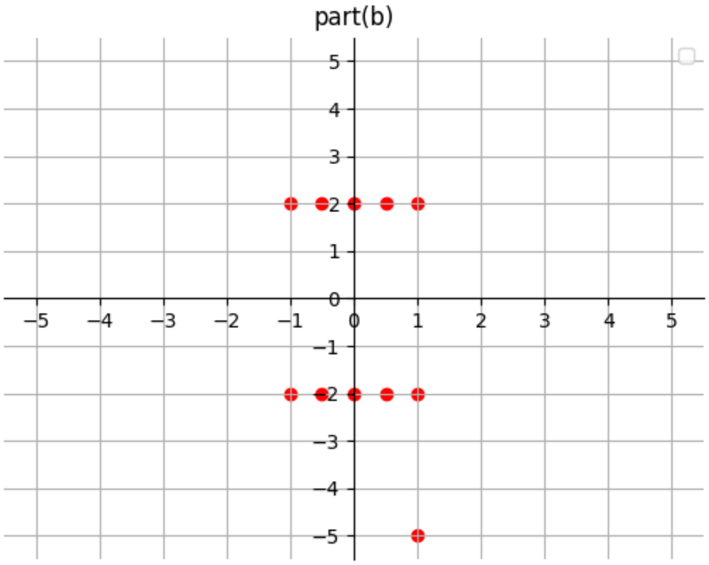


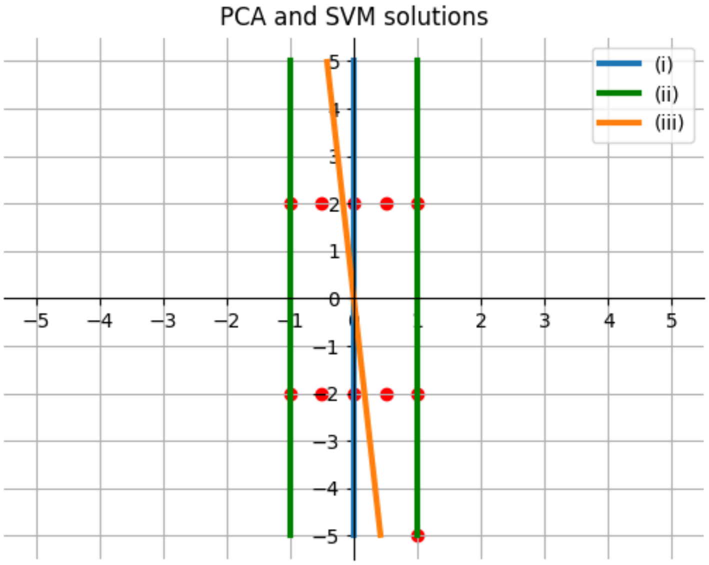
Solution: The key observation is that the our modified SVM finds a hyperplane that minimizes the maximum spread of the data along the orthogonal direction, whereas PCA finds a direction that maximizes the variance along that direction. An example of a dataset whether the two methods would yield the same solution is given below.



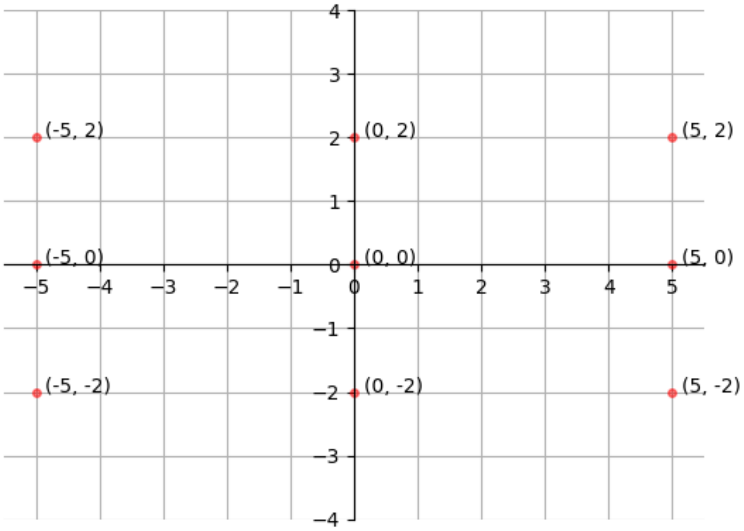


One way to make the solutions different is to add a point inside the margin that is not on the first principal axis, which does not change the solution of the modified SVM. On the other hand, the first principal component would lean towards the additional point to maximize the variance along the principal component & minimize reconstruction error.

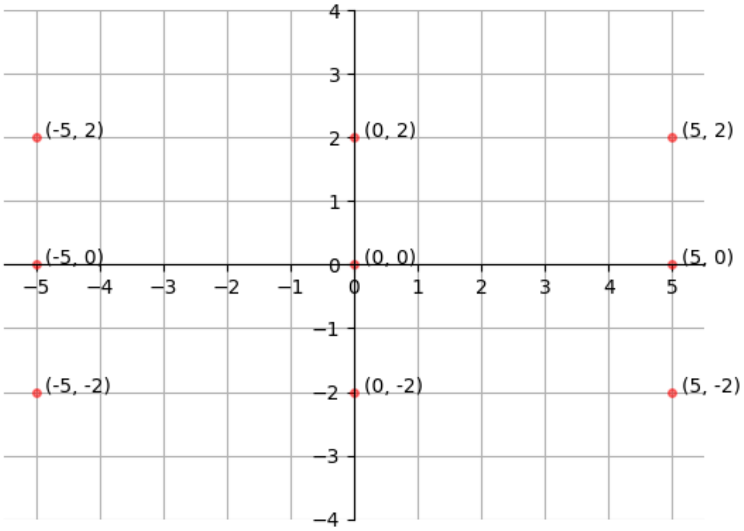




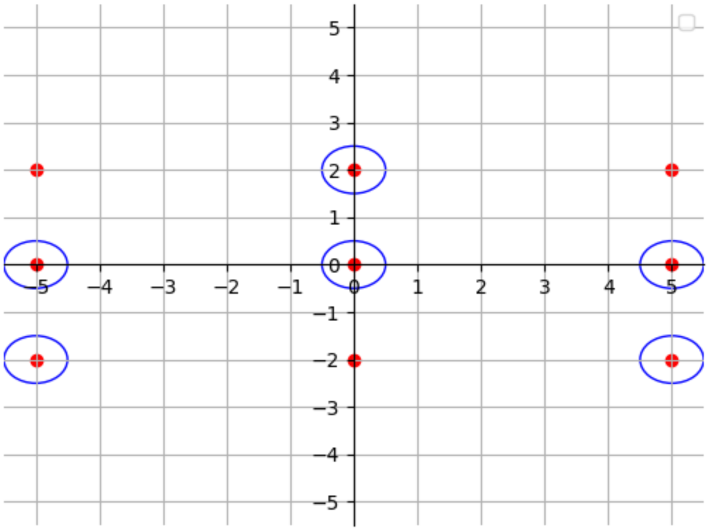
8. (4 pts) Consider the dataset shown in the plot below (note that this is **different** from the dataset in the previous part). **First, draw (i) the line such that $\mathbf{w}^{*\top}\mathbf{x}_{\text{test}} - b^* = 0$ and (ii) the two lines such that $\left(\mathbf{w}^{*\top}\mathbf{x}_{\text{test}} - b^*\right)/\|\mathbf{w}^*\|_2 = m^*$ and $\left(\mathbf{w}^{*\top}\mathbf{x}_{\text{test}} - b^*\right)/\|\mathbf{w}^*\|_2 = -m^*$, and label each with (i) or (ii). Then, find a largest set of data points you can remove such that the optimal solution $(\mathbf{w}^*, b^*, m^*)$ to the optimization problem in part 1 doesn't change.** Circle all points that are in this set.

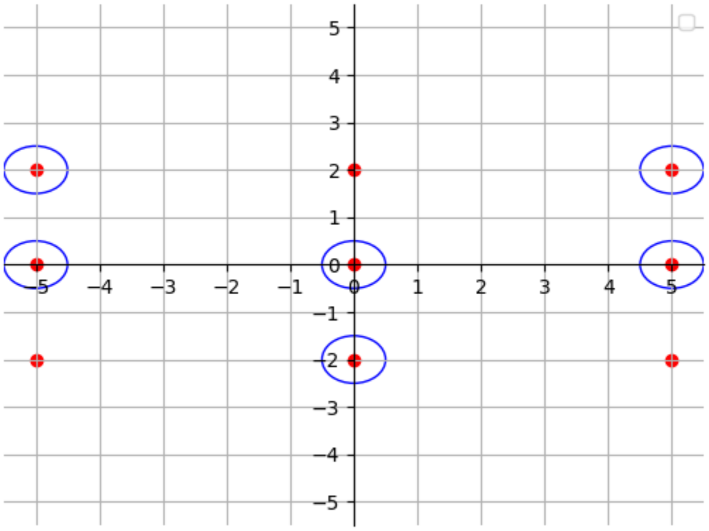


If you make a mistake, cross out the plot above and draw in the plot below:



Solution: It's not hard to figure out that the optimal hyperplane for the dataset is the horizontal axis, and the margin hyperplanes are a distance of 2 away from the hyperplane. Removing either set of points below wouldn't change the solution.





4 Multiple Choice (30 points)

For each of the following questions, bubble in *all* the correct answer choices. Each question may have more than one correct answer.

1. What of the following facts about maximum a posteriori (MAP) are true?

   □ Maximizes the probability of the parameters given the data

   □ Maximizes the probability of the data given the parameters

   □ Is equivalent to maximum likelihood if the prior on the parameters is uniform

   □ Is equivalent to maximum likelihood if the prior on the data is uniform

   □ None of the above

2. Which of the following regularization methods for linear regression is equivalent to imposing a $\mathrm{MultivariateNormal}(\mathbf{0}, \sigma^2\mathbf{I})$ prior on the parameters?

   □ Penalizing the ℓ_1-norm of the parameters

   □ Penalizing the squared ℓ_1-norm of the parameters

   □ Penalizing the ℓ_2-norm of the parameters

   □ Penalizing the squared ℓ_2-norm of the parameters

   □ None of the above

3. Which of the following properties must be true in a *linear* classifier for a binary classification problem?

   □ Predicted label (which is either 0 or 1) must be a linear function of the parameters

   □ Predicted label (which is either 0 or 1) must be a linear function of the data

   □ Decision boundary must be a linear function of the data

   □ Density of the class-conditional distribution must be a linear function of the data

   □ None of the above

4. What are some ways to reduce *overfitting* in a neural net classifier?

   □ Increase the number of hidden layers

   □ Force some elements of the weight matrix to have the same value

   □ Apply weight decay, i.e.: ℓ_2 regularization on the weights

   □ Reduce the number of training examples

   □ None of the above

5. Which of the following optimization algorithms may fail to find the global minimum of an objective function that is convex, but not necessarily differentiable everywhere?

   ☐ Setting the gradient to zero and solve

   ☐ Gradient descent

   ☐ Coordinate descent

   ☐ All of the above are guaranteed to find the global minimum

6. What are some ways to reduce *overfitting* in a decision tree classifier?

   ☐ Decrease the depth of the tree

   ☐ Decrease the total number of nodes in tree

   ☐ Decrease the number of training examples in each leaf node

   ☐ None of the above

7. How does increasing the number of nearest neighbors k in a k-nearest neighbors classifier affect the bias-variance tradeoff in classification error?

   ☐ Increases bias

   ☐ Decreases bias

   ☐ Increases variance

   ☐ Decreases variance

   ☐ None of the above

8. Let $\mathbf{x}$ be a random variable denoting the data, and y be a random variable denoting the class label. Which of the following distributions does a generative classification method fit models to?

   ☐ $p(y|\mathbf{x})$

   ☐ $p(\mathbf{x}|y)$

   ☐ $p(y)$

   ☐ $p(\mathbf{x})$

   ☐ None of the above

9. Which of the following properties of expectation-maximization (E-M) are true?

   ☐ Finds a parameter setting that is the global maximum of marginal log-likelihood of the data at convergence

   ☐ Finds a parameter setting that is the local maximum of marginal log-likelihood of the data at convergence

   ☐ Marginal log-likelihood of the data does not decrease from iteration to iteration

   ☐ Value of the variational lower bound/evidence lower bound (ELBO) does not decrease from iteration to iteration

   ☐ None of the above

10. Consider two versions of logistic regression, one with ℓ_2 regularization on the parameters, and one without any regularization on the parameters. How do they behave on a dataset with two linearly separable classes when trained using (non-stochastic) gradient descent?

    □ Always converges without regularization with a sufficiently small constant learning rate

    □ Never converges without regularization with any constant learning rate

    □ Always converges with regularization with a sufficiently small constant learning rate

    □ Never converges with regularization with any constant learning rate

    □ None of the above

Solution:

1. Answer: A, C.

   MAP maximizes the posterior probability of the parameters given the data (which is equivalent to MLE with a uniform prior on the parameters).

2. Answer: D.

   Penalizing the squared ℓ_2-norm. The squared ℓ_2-norm of the weight vector is equivalent to the log-density of the zero-mean isotropic Gaussian prior on the weights up to constant terms and factors. The MAP objective adds the log-density of the prior on the weights to the log-likelihood of the training data, which is equivalent to adding a penalty on the squared ℓ_2-norm.

3. Answer: C.

   The predicted label is binary and so cannot be a linear function of either the parameters or the data, since the output of a linear function must be continuous. So, in general, the predicted label of any classifier cannot be a linear function of the parameters or the data. The density of any distribution cannot be a linear function, since the integrating it over $\mathbb{R}^d$ would result in ∞, which implies that it would not be a valid probability distribution. A linear classifier refers to a classifier with a linear decision boundary.

4. Answer: B, C.

   Forcing some elements to have the same value (one example of this being convolutional architectures) or using ℓ_2 regularization reduces model capacity and therefore reduces overfitting. The other examples in fact increase overfitting.

5. Answer: A, B, C.

   At the minimum, the gradient is either zero or does not exist. So, setting the gradient to zero may not find the minimum. When running gradient descent, if the gradient at current iterate does not exist, the behaviour is undefined. So, gradient descent may fail. Coordinate descent requires both convexity and differentiability to converge to the global minimum. For an example of a convex function where coordinate descent fails, see page 12 of lecture note 20.

6. Answer: A, B.

   Restrict depth, total number of nodes. Both of these methods reduce the capacity (i.e., complexity, or expressiveness) of the model, which reduces its ability to overfit to the training data. Placing an *upper* bound on the number of training examples in a leaf node has the opposite effect: it forces the learning algorithm to split nodes that contain too many training examples, which leads to the creation of additional nodes and makes the model more complex.

7. Answer: A, D.

   Increases bias, decreases variance. Increasing k leads to smoother decision boundaries.

8. Answer: B, C.

   Generative classification methods, like LDA or QDA, fit models to the prior over labels, $p(y)$, and the likelihood/class-conditional distribution, $p(\mathbf{x}|y)$. Even though other distributions like $p(y|\mathbf{x})$ and $p(\mathbf{x})$ can be computed using Bayes' rule, we do not fit models to these distributions. Contrast this with discriminative classification methods, like logistic regression or probit, which fit models to the posterior, $p(y|\mathbf{x})$.

9. Answer: C, D. The variational lower bound/ELBO $\mathcal{L}(q, \theta)$ may not be differentiable w.r.t. the parameters of the joint distribution of the observed and latent variables $p(\mathbf{x}, \mathbf{z}; \theta)$ or the parameters of the averaging distribution $q(\mathbf{z}|\mathbf{x})$. Even when $\mathcal{L}(q, \theta)$ is convex in q and θ and $\max_q \mathcal{L}(q, \theta)$ is convex in θ, E-M, which is coordinate ascent on the variational lower bound, may fail to converge to the global maximum of the variational lower bound, which is also the global maximum of the marginal log-likelihood. (For an example of such a function, see page 12 of lecture note 20.) Since $\max_q \mathcal{L}(q, \theta)$ is the marginal log-likelihood, in this case, the only local maximum is the global maximum, and so E-M does not converge to a local maximum either. As shown in lecture note 20, the value of $\mathcal{L}(q, \theta)$ never decreases over time. After each E-step, the value of $\mathcal{L}(q, \theta)$ is the marginal log-likelihood at θ, and since the value of $\mathcal{L}(q, \theta)$ never decreases over time, the value $\mathcal{L}(q, \theta)$ after each E-step never decreases over time, and so marginal log-likelihood never decreases over time.

10. Answer: B, C.

    Logistic regression does not converge without regularization, since we can keep increasing the norm of the weight vector to make the decision boundary sharper and maximize the log-likelihood of the linearly-separable training data. On the other hand, it does converge with regularization, since penalizing the Euclidean norm of the weights constrains the weight vector to lie within a ball with a finite radius.

The PDF of a multivariate n-dimensional Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is given by

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Doodle page! Draw us something if you want or give us suggestions or complaints. You can also use this page to report anything suspicious that you might have noticed.