1. (2 points) Suppose you have a function

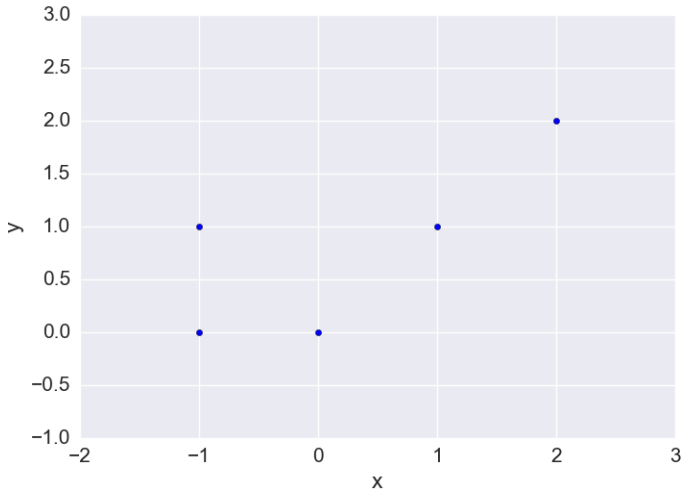
$$\varphi(c) = \arg\min_{w \in \mathbf{R}^d} \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i[w^T x_i]\right) + \frac{1}{2}||w||^2.$$

where $c > 0$. Show how we could use this $\varphi(c)$ to find a minimizer $w^* \in \mathbf{R}^d$ of the following objective function (where $\lambda > 0$):

$$J(w) = \frac{1}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i[w^T x_i]\right) + \lambda||w||^2$$

1/2lambda J(w)
c = 1/2 lambda

2. Let $\mathcal{X} = \mathcal{Y} = \mathcal{A} = \mathbb{R}$. Suppose you receive the (x, y) data points (-1,1), (-1,0), (0,0), (1,1), and (2,2).



(a) Assume we're using the 0-1 loss function $\ell(a, y) = \mathbf{1}(a \neq y)$.

i. (1 point) Suppose we restrict to the hypothesis space $\mathcal{F}_c$ of constant functions. Give an empirical risk minimizer $\hat{f}(x)$.

f(x) = 0 or f(x) = 1
then R(f) = 3/5

ii. (1 point) Suppose we restrict to the hypothesis space $\mathcal{F}_c$ of constant functions. What is $\hat{R}(\hat{f})$, the empirical risk of $\hat{f}$, where $\hat{f}$ is an empirical risk minimizer.

iii. (1 point) Suppose we restrict to the hypothesis space $\mathcal{F}_\ell$ of linear functions. Give an empirical risk minimizer $\hat{f}(x)$.

f(x) = x

(b) Now assume we're using the absolute loss function $\ell(a, y) = |a - y|$.

i. (1 point) What is the minimum empirical risk achievable over the hypothesis space of all functions?

y = 0 xin [-1,0]
y = x x >0
risk is 1/5

3. (a) (1 point) Consider the following version of the elastic-net objective:

$$J(w) = \frac{1}{n}\|Xw - y\|_2^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2.$$

Our training data is $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, sampled i.i.d. from some distribution P. As usual, the design matrix $X \in \mathbb{R}^{n \times d}$ has x_i as its ith row, and $y \in \mathbb{R}^n$ has y_i as its i'th coordinate. Which ONE of the following hyperparameter settings is most likely to give a sparse solution?

☐ $\lambda_1 = 0, \lambda_2 = 1$ ☐ $\lambda_1 = 1, \lambda_2 = 0$ ☐ $\lambda_1 = 0, \lambda_2 = 0$

correct

4. The penalized empirical risk for $f \in \mathcal{F}$ and dataset $\mathcal{D}$ is given by

$$J(f; \mathcal{D}) = \hat{R}(f; \mathcal{D}) + \lambda \Omega(f),$$

regular lasso to avoid overfitting

where $\Omega : \mathcal{F} \to [0, \infty)$ is a regularization function, $\lambda > 0$ is a regularization parameter, and

$$\hat{R}(f; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} \ell(f(x), y)$$

EMR

is the empirical risk of f for the data $\mathcal{D}$, where $|\mathcal{D}|$ is the size of the set $\mathcal{D}$.

(a) (1 point) Suppose we use an iterative descent method to minimize $J(f; \mathcal{D})$ for some training data $\mathcal{D}$. Let $f^{(i)}$ be the prediction function at the i'th iteration. If our goal is to find a minimizer $\hat{f} \in \arg\min_{f \in \mathcal{F}} J(f; \mathcal{D})$, which ONE of the following is the better stopping condition?

we don't care R

- ☐ $\hat{R}(f^{(i)}, \mathcal{D}) - \hat{R}(f^{(i+1)}, \mathcal{D}) < \epsilon$, for some appropriately chosen $\epsilon > 0$.
- correct ☐ $J(f^{(i)}, \mathcal{D}) - J(f^{(i+1)}, \mathcal{D}) < \epsilon$, for some appropriately chosen $\epsilon > 0$.

(b) (1 point) A friend reminds us that our real goal is to find an f that has small risk, i.e. a small value of $R(f) = \mathbb{E}\ell(f(x), y)$. Suppose we have found $\tilde{f}$ using $\mathcal{D}$ and we have an independent validation set $\mathcal{D}_{\text{val}}$ from the same distribution as $\mathcal{D}$. Select ALL of the following that are unbiased estimators of $R(\tilde{f})$:

using training data to mini R

☐ $J(\tilde{f}, \mathcal{D}_{val})$ ☐ $\hat{R}(\tilde{f}, \mathcal{D})$ ☐ $J(\tilde{f}, \mathcal{D})$ ☐ $\hat{R}(\tilde{f}, \mathcal{D}_{val})$

we don't care J this time

correct

(c) (2 points) A friend thinks that we are running too many iterations of the optimization procedure to minimize $J(f; \mathcal{D})$, when our real goal is to find f with small risk. Let

$$f^* \in \arg\min_f \mathbb{E}\ell(f(x), y)$$

$$f_{\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(x), y)$$

don't depend on data as well

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} J(f; \mathcal{D})$$

which depend on data

and again let $f^{(i)}$ be the prediction function at the current iteration. Select ALL of the following that would support your friend's claim that it's time to stop the optimization algorithm:

- ☐ $J(f^{(i)}, \mathcal{D}) = J(\hat{f}, \mathcal{D})$
- ☐ $R(f^{(i)}) < R(\hat{f})$ get close to f^ correct
- ☐ $\hat{R}(f^{(i)}, \mathcal{D}) < \hat{R}(\hat{f}, \mathcal{D})$ maybe not need J here, EMR maybe greater than the R^, see the photo picture
- ☐ $R(\hat{f}) - R(f_{\mathcal{F}})$ is significantly smaller than $R(f_{\mathcal{F}}) - R(f^*)$ no relationship, no meaning

(d) (1 point) Your friend suggests using your validation data $\mathcal{D}_{val}$ for "early stopping." Which ONE of the following is the BEST suggestion for an early stopping rule that **you could use in practice**:

find the minimize of J not R

- ☐ $\hat{R}(f^{(i-100)}, \mathcal{D}_{val}) - \hat{R}(f^{(i)}, \mathcal{D}_{val}) < \epsilon$ for some appropriately chosen $\epsilon > 0$.
- ☐ $J(f^{(i-100)}, \mathcal{D}_{val}) - J(f^{(i)}, \mathcal{D}_{val}) < \epsilon$ for some appropriately chosen $\epsilon > 0$.
- ☐ $\hat{R}(f^{(i)}, \mathcal{D}_{val}) - \hat{R}(\hat{f}, \mathcal{D}_{val}) < \epsilon$ for some appropriately chosen $\epsilon > 0$.

answer: A

in practice, we don't have f^

- ☐ $J(f^{(i)}, \mathcal{D}_{val}) - J(\hat{f}, \mathcal{D}_{val}) < \epsilon$ for some appropriately chosen $\epsilon > 0$.

5. Decide whether the following statements apply to full batch gradient descent (GD), mini-batch GD, neither, or both. Assume we're minimizing a differentiable, convex objective function $J(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$, and we are currently at w_t, which is not a minimum. For full batch GD, take $v = \nabla_w J(w_t)$, and for minibatch GD take v to be a minibatch estimate of $\nabla_w J(w_t)$ based on a random sample of the training data.

   (a) (1 point) For any step size $\eta > 0$, after applying the update rule $w_{t+1} \leftarrow w_t - \eta v$, we must have $J(w_{t+1}) < J(w_t)$. (Choose ONE answer below.)
   ☐ Full batch ☐ Minibatch ☐ Both ☐ Neither
   correct

   (b) (1 point) There must exist some $\eta > 0$ such that after applying the update rule $w_{t+1} \leftarrow w_t - \eta v$ we have $J(w_{t+1}) < J(w_t)$. (Choose ONE answer below.)
   ☐ Full batch ☐ Minibatch ☐ Both ☐ Neither
   correct

   (c) (1 point) v is an unbiased estimator of the full batch gradient. (Choose ONE answer below.)
   ☐ Full batch ☐ Minibatch ☐ Both ☐ Neither
   correct

$\eta > 0$. Circle all statements below that we know will be true for small enough $\eta > 0$ (circling none of them is allowed):

(a) $J(w_{i+1}) < J(w_i)$

(b) $\|w^* - w_{i+1}\| < \|w^* - w_i\|$

SOLUTION: (a) is true by definition of gradient and (b) is true because a subgradient step takes us closer to the minimizer (proved in slides).

2. [1] $J(w) : \mathbf{R}^d \to \mathbf{R}$ is a **convex** objective function with minimizer w^*. Assume w_i is not a minimizer of $J(w)$. Let $g \in \partial J(w_i)$ be a **subgradient** of J at w_i, and let $w_{i+1} = w_i - \eta g$, for some $\eta > 0$. Circle all statements below that we know will be true for small enough $\eta > 0$ (circling none of them is allowed):

(a) $J(w_{i+1}) < J(w_i)$

(b) $\|w^* - w_{i+1}\| < \|w^* - w_i\|$

SOLUTION: Just (b), by reason above. Subgradient step doesn't necessarily decrease objective function value (also discussed in slides).

3. [1] Let $J(w) = \frac{1}{n}\sum_{i=1}^{n} J_i(w)$, where each $J_i(w) : \mathbf{R}^d \to \mathbf{R}$ is **convex and differentiable.** Suppose $J(w)$ has minimizer w^*. Let $g = \nabla J_1(w)$. [Please **note** the subscript on J.] Let $w_{i+1} = w_i - \eta g$, for some $\eta > 0$. Circle all statements below that we know will be true for small enough $\eta > 0$ (circling none of them is allowed):

(a) $J(w_{n+1}) < J(w_n)$

(b) $\|w^* - w_{n+1}\| < \|w^* - w_n\|$

SOLUTION: Neither. This is a stochastic gradient step, for which we have no guarantee about the change of any individual step. Over the long term, SGD eventually takes us to the minimizer under some conditions.

7 Perceptron

The **perceptron loss** is given by

$$\ell(\hat{y}, y) = \max\{0, -\hat{y}y\}.$$

And consider the hypothesis space of linear functions $\mathcal{H} = \left\{f \mid f(x) = w^T x,\, w \in \mathbf{R}^d\right\}$.

1. [1] Is the perceptron loss a margin-based loss? Justify your answer.
SOLUTION: A margin-based loss is a loss function that depends on y and $\hat{y}$ only via the "margin", which is the product $y\hat{y}$. ℓ is clearly a margin loss.

2. [1] Suppose we have a linear function $f(x) = w^T x$, for some $w \in \mathbf{R}^d$. Geometrically, we say that the hyperplane $H = \{x \mid f(x) = 0\}$ separates the dataset $\mathcal{D} = ((x_1, y_1), \ldots, (x_n, y_n)) \in \mathbf{R}^d \times \{-1, 1\}$ if all x_i corresponding to $y_i = -1$ are strictly on one side of H, and all x_i corresponding to $y_i = 1$ are strictly on the other side of H. ("Strictly" here means that no x_i's lie on H.) Give a mathematical formulation of the necessary and sufficient conditions for $f(x) = w^T x$ to separate $\mathcal{D}$. [Hint: Answer will involve the data points and the function f.]
SOLUTION:

$$y_i f(x_i) > 0 \; \forall i \in \{1, \ldots, n\}$$

3. [1] In the homework we showed that if our prediction function $f(x) = w^T x$ separates a dataset $\mathcal{D}$, then the total perceptron loss on $\mathcal{D}$ is 0. The converse is not true: we may have total perceptron loss 0, but $f(x)$ may not separate $\mathcal{D}$. Explain how this can happen.
SOLUTION: We may have this if any x_i lies on the hyperplane — i.e. if $f(x_i) = w^T x_i = 0$. When this happens, the loss on this example will be 0, but the point is not strictly on the correct side of the hyperplane.

8 Regularized Perceptron

Consider a hypothesis space of linear functions $\mathcal{H} = \left\{ f \mid f(x) = w^T x,\, w \in \mathbf{R}^d \right\}$. Let $\ell(\hat{y}, y) = \max\{0, -\hat{y}y\}$ be the Perceptron loss. Consider the objective function

$$J(w) = \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \max\left\{0, -y_i w^T x_i\right\}.$$

We are interested in finding the minimizer of $J(w)$ **subject to** the constraint that $\|w\|^2 \geq 1$.

1. [2] Let $J_1(w; x, y) = \frac{1}{2}\|w\|^2 + c\max\left\{0, -yw^T x\right\}$. Give a subgradient g of $J_1(w; x, y)$ with respect to w. The subgradient will be a function of x, y, c, and w.
SOLUTION:

$$g \quad = \quad \begin{cases} -cyx + w & \text{for } yw^T x < 0 \\ w & \text{for } yw^T x \geq 0. \end{cases}$$

2. [1] Write the Lagrangian for the problem of minimizing $J(w)$ **subject to** the constraint that $\|w\|^2 \geq 1$.
SOLUTION:

$$L(w, \lambda) \quad = \quad J(w) + \lambda\left(1 - \|w\|^2\right)$$

3. [2] Assuming it's attained, give an expression for the [primal] optimal value of the optimization problem in terms of the Lagrangian. **Explain** why this gives the same optimal value as the original problem.
SOLUTION: The primal optimal value is

$$p^* \quad = \quad \min_w \sup_{\lambda \geq 0} L(w, \lambda)$$

for the following reason: If w is feasible, then the inner supremum is just $J(w)$, and otherise it's ∞. The outer minimum will only ever select w for which the inner optimization is $J(w)$. So it's equivalent to the original problem.

4. [1] State the dual objective function and the dual optimization problem in terms of the Lagrangian function.
SOLUTION: The dual objective function is

$$g(\lambda) \quad = \quad \inf_w L(w, \lambda),$$

DSGA 1003: Test #1 Practice Problems

Part I

From 2015 Exam

1 True / False Questions

1. (**True or False**, 1 pt) When using (unregularized) linear regression, adding new features always improves the performance on training data, or at least never make it worse.

2. (**True or False**, 1 pt) When using a (unregularized) linear regression, adding new features always improves the performance on test data, or at least never make it worse.

3. (**True or False**, 1 pt) Overfitting is more likely when the set of training data is small.

4. (**True or False**, 1 pt) Overfitting is more likely when the hypothesis space is small.

5. (**True or False**, 1 pt) Approximation error decreases to zero as the amount of training data goes to infinity.

6. (**True or False**, 1 pt) Suppose we fit Lasso regression to a data set. If we rescale one of the features by multiplying it by 10, and we then refit Lasso regression with the same regularization parameter, then it is more likely for that feature to be excluded from the model. [*NOTE: This question is not ideal because we haven't done any rigorous proof for this. But I think one direction is clearly better.*]
   Solutions: 1) True 2) False 3) True 4) False 5) False 6) False

2 Short Answer

1. (1 pt) Circle all of the loss functions that may lead to sparse support vectors: **hinge loss, squared hinge loss, logistic loss, square loss**. (*Hint: Consider the homework problem in which you characterize the support vectors in the SVM or Perceptron solution in terms of what happens during SGD.*)

   Solution: With the hinge loss-based loss functions, we have no loss when the margin exceeds 1, while for the other losses, we have a loss no matter how big the margin. We cannot prove that the hinge losses will give us sparsity because they may not, but they often do give sparsity of support vectors. We can definitely prove that logistic loss and square loss do not have any sparsity. We can see this simply from the SGD update. No matter what w we start at,

every example (x_i, y_i) will trigger an update, and thus x_i enters the linear combination for the expression for w.

2. (4 pts) We have a dataset $\mathcal{D} = \{(0,1), (1,4), (2,3)\}$ that we fit by minimizing an objective function of the form:

$$J(\alpha_0, \alpha_1) = \frac{1}{3}\sum_{i=1}^{3}(\alpha_0 + \alpha_1 x_i - y_i)^2 + \lambda_1(|\alpha_0| + |\alpha_1|) + \lambda_2(\alpha_0^2 + \alpha_1^2),$$

and the corresponding fitted function is given by $f(x) = \alpha_0 + \alpha_1 x$. We tried four different settings of λ_1 and λ_2, and the results are shown in Figure 1. For each of the following parameter settings, give the number of the plot that shows the resulting fit.

   (a) (1 pt) $\lambda_1 = 0$ and $\lambda_2 = 0$.
   (b) (1 pt) $\lambda_1 = 5$ and $\lambda_2 = 0$.
   (c) (1 pt) $\lambda_1 = 0$ and $\lambda_2 = 10$.
   (d) (1 pt) $\lambda_1 = 0$ and $\lambda_2 = 2$.

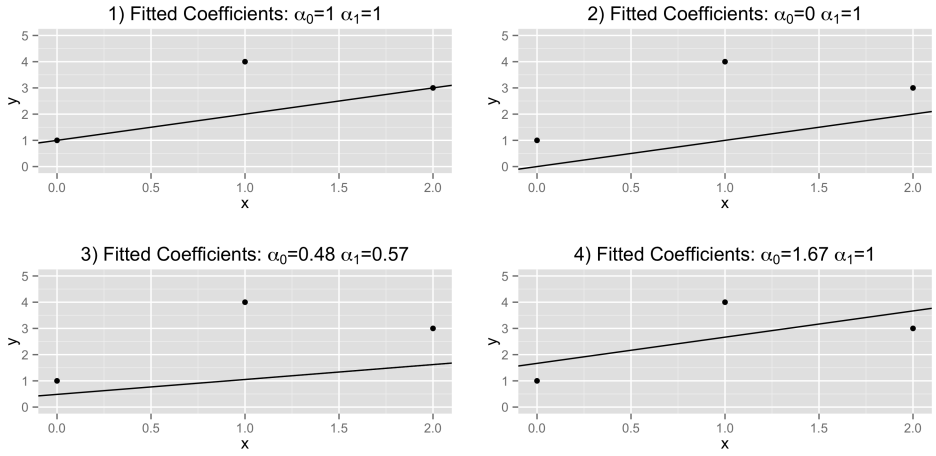


Figure 1: Linear fits with different penalizations.

Solution: a) 4 (fits the data the best) b) 2 (sparse fit is a good hint that it's L1, and then we're sure by process of elimination) c) 3 (more regularized than 1, smallest L2 norm) d) 1

3. (2 pts) Show that the following kernel function is a Mercer kernel (i.e. it represents an inner product):

$$k(x, y) = \frac{x^T y}{\|x\|\|y\|},$$

where $x, y \in \mathbf{R}^d$.

Solution: For $\phi(x) = \frac{x}{\|x\|}$, we have

$$k(x, y) = \langle \phi(x), \phi(y) \rangle .$$

4. (2 pts) Consider the binary classification problem shown in Figure 2: Denote the input space

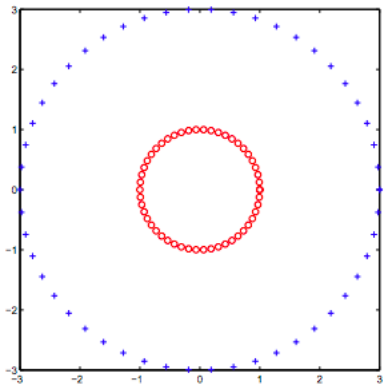


Figure 2: For a short-answer problem.

by $\mathcal{X} = \{(x_1, x_2) \in \mathbf{R}^2\}$. Give a feature mapping for which a linear classifier could perfectly separate the two classes shown.

Solution: $(x_1, x_2) \mapsto (1, x_1, x_2, x_1^2, x_2^2)$ works... Anything that allows you to construct $f(x) = ax_1^2 + ax_2^2$ as a linear combination of the feature vector would work.

3 Hypothesis Spaces

1. (2 pt) For the input space $\mathcal{X} = \mathbf{R}$, consider the following two hypothesis spaces:

$$\mathcal{F}_1 = \{f(x) = e^{w_1}x + w_2 x \mid w_1, w_2 \in \mathbf{R}\} \qquad \mathcal{F}_2 = \{f(x) = wx \mid w \in \mathbf{R}\}$$

Suppose we are selecting hypotheses using empirical risk minimization (without any penalty). Are there any situations in which one of these hypothesis spaces would be preferred to the other? Why?

Solution: The two hypothesis spaces are the same. So either 1) no preference, or 2) prefer $\mathcal{F}_2$ because there's no annoying unidentifiability that $\mathcal{F}_1$ has (in other words, in $\mathcal{F}_1$ there are multiple parameter settings that give the same prediction functions, while not the case for $\mathcal{F}_2$).

2. (2 pt) Same question, with the following hypothesis spaces:

$$\mathcal{F}_1 = \{f(x) = e^{w_1}x \mid w_1 \in \mathbf{R}\} \qquad \mathcal{F}_2 = \{f(x) = wx \mid w \in \mathbf{R}\}$$

Solution: The hypothesis spaces are different. $\mathcal{F}_1$ makes sense if you know your prediction function should always output something that's the same sign as x. Otherwise, $\mathcal{F}_2$.

1. (7 points) Consider a binary classification problem. For class $y = 0$, x is sampled from $\{1, 2, 3, 4, 5, 6, 7, 8\}$ with equal probability; for class $y = 1$, x is sampled from $\{7, 8, 9, 10\}$ with equal probability. Assume that both classes are equally likely. Let $f^* : \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \to \{0, 1\}$ represent the Bayes prediction function for the given setting under $0 - 1$ loss. Find f^* and calculate the Bayes risk.

Solution: 0-1 loss:

$$l(a, y) = 1(a \neq y) := \begin{cases} 1 & \text{if } a \neq y \\ 0 & \text{otherwise.} \end{cases}$$

Risk:

$$\begin{aligned} R(f) &= \mathbb{E}\left[1(f(x) \neq y)\right] \quad = 0 \cdot \mathbb{P}\left(f(x) = y\right) + 1 \cdot \mathbb{P}\left(f(x) \neq y\right) \\ &= \mathbb{P}\left(f(x) \neq y\right), \end{aligned}$$

which is just the misclassification error rate.

Bayes prediction function is just the assignment to the most likely class,

$$f^*(x) = \underset{c \in \{0,1\}}{argmax}\ p(y = c|x)$$

Therefore:

$$f^*(x) = \begin{cases} 0 & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 1 & \text{if } x \in \{7, 8, 9, 10\} \end{cases}$$

Under $0 - 1$ loss, risk is the probability of mis-classification. $f^*(x)$ mis-classifies points from class 0 occurring in $\{7, 8\}$ as class 1. Hence, bayes risk is

$$\begin{aligned} p(y = 0, x \in \{7, 8\}) &= p(x \in \{7, 8\}|y = 0)p(y = 0) \\ &= \frac{1}{4} \times \frac{1}{2} \\ &= \frac{1}{8} \end{aligned}$$

2. Consider the statistical learning problem for the distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \mathcal{Y} = \mathbf{R}$. A labeled example $(x, y) \in \mathbf{R}^2$ sampled from $\mathcal{D}$ has probability distribution given by $x \sim \mathcal{N}(0, 1)$ and $y|x \sim \mathcal{N}(f^*(x), .1)$, where $f^*(x) = \sum_{i=0}^{5}(i+1)x^i$.

   Let P_k denote the set of all polynomials of degree k on $\mathbf{R}$–that is, the set of all functions of the form $f(x) = \sum_{i=0}^{k} a_i x^i$ for some $a_1, \ldots, a_k \in \mathbf{R}$.

   Let D_m be a training set $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbf{R} \times \mathbf{R}$ drawn i.i.d. from $\mathcal{D}$. We perform empirical risk minimization over a hypothesis space $\mathcal{H}$ for the square loss. That is, we try to find $f \in \mathcal{H}$ minimizing

$$\hat{R}_m(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x) - y)^2$$

   (a) (2 points) If we change the hypothesis space $\mathcal{H}$ from $P_3(x)$ to $P_4(x)$ while keeping the same training set, select **ALL** of the following that **MUST** be true:

   - ☐ Approximation error increases or stays the same.
   - ■ **Approximation error decreases or stays the same.**
   - ☐ Estimation error increases or stays the same.
   - ☐ Bayes risk decreases.

   (b) (2 points) If we change the hypothesis space $\mathcal{H}$ from $P_5(x)$ to $P_6(x)$ while keeping the same training set, select **ALL** of the following that **MUST** be true:

   - ■ **Approximation error stays the same.**
   - ☐ Estimation error stays the same.
   - ☐ Optimization error stays the same.
   - ■ **Bayes risk stays the same.**

   (c) (2 points) If we increase the size of the training set m from 1000 to 5000 while keeping the same hypothesis space $P_5(x)$, select **ALL** of the following that **MUST** be true:

   - ■ **Approximation error stays the same.**
   - ☐ Estimation error decreases or stays the same.
   - ■ **The variance of $\hat{R}_m(f)$ for $f(x) = x^2$ decreases.**
   - ■ **Bayes risk stays the same.**

3. Let D_t denote a training set $(x_1, y_1), \ldots, (x_{n_t}, y_{n_t}) \in \mathbf{R}^d \times \{-1, 1\}$ and D_v a validation set $(x_1, y_1), \ldots, (x_{n_v}, y_{n_v}) \in \mathbf{R}^d \times \{-1, 1\}$. The training set D_t is linearly separable. Define $J(\theta) = \frac{1}{n_t} \sum_{(x,y) \in D_t} \ell(m)$, where $\ell(m)$ is a margin-based loss function, and m is the margin defined by $m = y(\theta^T x)$.

   We have run an iterative optimization algorithm for 100 steps and attained $\tilde{\theta}$ as our approximate minimizer of $J(\theta)$.

   Denote the training accuracy by $\alpha(D_t) = \frac{1}{n_t} \sum_{(x,y) \in D_t} \mathbf{1}(y\tilde{\theta}^T x > 0)$ and the validation accuracy by $\alpha(D_v) = \frac{1}{n_v} \sum_{(x,y) \in D_v} \mathbf{1}(y\tilde{\theta}^T x > 0)$.

   (a) Answer the following for the logistic loss $\ell(m) = \log(1 + e^{-m})$:

      i. (1 point) **F** **True or False**: Achieving 100% training accuracy ($\alpha(D_t) = 1$) implies that we have achieved a minimizer of the objective function ($\tilde{\theta} \in \arg\min_\theta J(\theta)$).

      ii. (1 point) **F** **True or False**: Achieving 100% **validation** accuracy ($\alpha(D_v) = 1$) implies that we have achieved a minimizer of the objective function ($\theta_t \in \arg\min_\theta J(\theta)$).

   (b) Answer the following for the hinge loss $\ell(m) = \max(0, 1 - m)$:

      i. (1 point) **F** **True or False**: Achieving 100% training accuracy ($\alpha(D_t) = 1$) implies that we have achieved a minimizer of the objective function ($\tilde{\theta} \in \arg\min_\theta J(\theta)$).

      ii. (1 point) **T** **True or False**: Achieving a minimizer of the objective function ($\tilde{\theta} \in \arg\min_\theta J(\theta)$) implies we have achieved **training** accuracy 100% ($\alpha(D_t) = 1$).

   (c) Answer the following for the perceptron loss $\ell(m) = \max(0, -m)$:

      i. (1 point) **T** **True or False**: Achieving 100% training accuracy ($\alpha(D_t) = 1$) implies that we have achieved a minimizer of the objective function ($\tilde{\theta} \in \arg\min_\theta J(\theta)$).

      ii. (1 point) **F** **True or False**: Achieving a minimizer of the objective function ($\tilde{\theta} \in \arg\min_\theta J(\theta)$) implies we have achieved **training** accuracy 100% ($\alpha(D_t) = 1$).