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 $\lambda_2 = 2$.

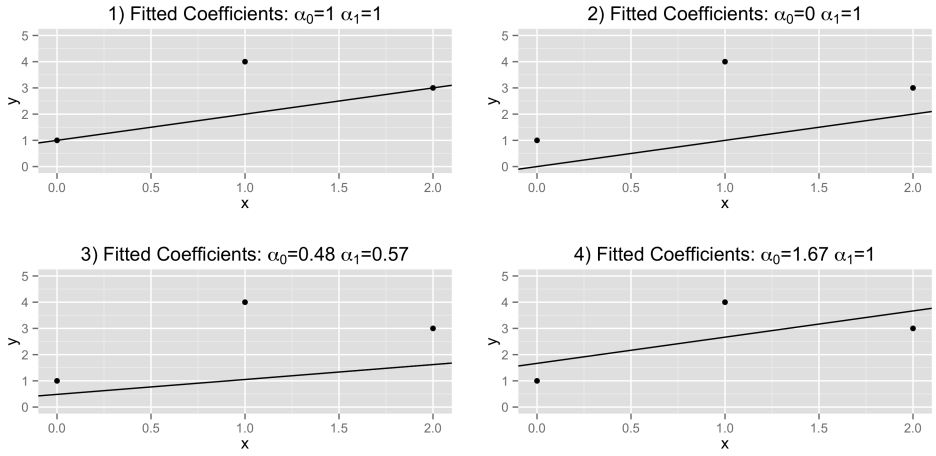


Figure 1: Linear fits with different penalizations.

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

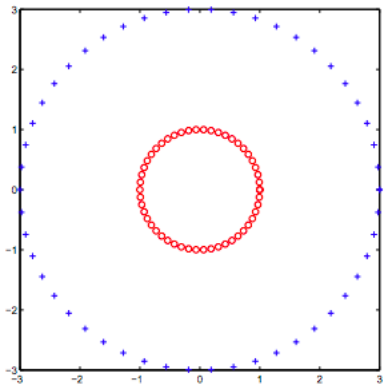


Figure 2: For a short-answer problem.

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

Solution: $(x_1, x_2) \mapsto \left(1, x_1, x_2, x_1^2, x_2^2\right)$ 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, $\mathcal{F}_2$.

4 Kernelizing Ridge Regression

Suppose our input space is $\mathcal{X}$=$\mathbf{R}^d$ and our output space is $\mathcal{Y} = \mathbf{R}$. Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training set from $\mathcal{X} \times \mathcal{Y}$. We'll use the "design matrix" $X \in \mathbf{R}^{n \times d}$, which has the input vectors as rows:

$$X = \begin{pmatrix} -x_1- \\ \vdots \\ -x_n- \end{pmatrix}.$$

Recall the ridge regression objective function:

$$J(w) = ||Xw - y||^2 + \lambda ||w||^2,$$

for $\lambda > 0$.

1. (4 pts) Derive a closed form expression for the minimizer of $J(w)$. .

   Solution:

$$\begin{array}{rcl} J(w) & = & (Xw-y)^T (Xw-y) + \lambda w^T w \\ \partial_w J(w) & = & 2X^T (Xw-y) + 2\lambda w \\ \partial_w J(w) = 0 & \iff & 2X^T Xw + 2\lambda w - 2X^T y = 0 \\ & \iff & (X^T X + \lambda I)w = X^T y \\ & \iff & w = (X^T X + \lambda I)^{-1} X^T y \end{array}$$

2. Using the representer theorem, give a kernelized version of the ridge regression objective function in terms of the Gram matrix (or the "kernel matrix") $K = XX^T$.

   Solution: By the representer theorem, $w = X^T \alpha$ for some $\alpha \in \mathbf{R}^n$. So we can write

$$\begin{array}{rcl} J(\alpha) & = & ||XX^T \alpha - y||^2 + \lambda ||X^T \alpha||^2 \\ & = & \|K\alpha - y\|^2 + \lambda \alpha^T K \alpha \end{array}$$

3. Solve for the minimizer of the kernelized ridge regression objective:

   Solution: (Throughout, using the fact that $K = K^T$)

$$\begin{array}{rcl} J(\alpha) & = & \alpha^T K^T K \alpha - 2\alpha^T K^T y + y^T y + \lambda \alpha^T K \alpha \\ \nabla J(\alpha) & = & 2K^2 \alpha - 2Ky + 2\lambda K \alpha \\ \nabla J(\alpha) = 0 & \iff & (K^2 + \lambda K)\alpha = Ky \\ & \iff & \alpha = (K^2 + \lambda K)^{-1} Ky \end{array}$$

4. (2 pts) Give a kernelized expression for Xw, the predicted values on the training points.

   Solution: Predictions are

$$Xw = XX^T \alpha = K\alpha = K(K^2 + \lambda K)^{-1} Ky$$

   (Note: This could be simplified, but one has to do it carefully, because we don't assume that K is invertible.

5. (1 pt) Give a kernelized expression for the prediction on new points, stored as rows in the matrix X_P.

   Solution: Predictions are

$$X_P w \quad = \quad X_P X^T \alpha$$

   This is already kernelized, since $X_P X^T$ is a matrix of inner products between x's, and α has already been kernelized.

5 Square Hinge Loss and Huberized Square Hinge Loss

The squared hinge loss is a margin loss given by

$$\ell(m) = \left[(1-m)_+\right]^2,$$

where $(m)_+ = m1(m > 0)$ is the "positive part" of m.

1. (2 pts) Suppose we have a training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_i \in \mathcal{X} = \mathbf{R}^d$ and $y_i \in \mathcal{Y} = \{-1, 1\}$, for all $i = 1, \ldots, n$. Consider the linear hypothesis space $\mathcal{F} = \left\{f(x) = w^T x \mid w \in \mathbf{R}^d\right\}$. Write the objective function $J(w)$ for ℓ_2-regularized empirical risk minimization with the square hinge loss over the space $\mathcal{F}$, where $\mathcal{F}$ is parameterized by w.

   Solution:

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

   for $\lambda > 0$.

2. (2 pts) It turns out that $J(w)$ is differentiable at every w. Give the derivative of $J(w)$.

   Solution:

$$\frac{\partial J(w)}{\partial w} \quad = \quad 2\lambda w + \frac{1}{n}\sum_{i=1}^{n}\begin{cases} -2\left(1 - y_i w^T x_i\right) y_i x_i & y_i w^T x < 1 \\ 0 & \text{otherwise} \end{cases}$$

3. (3 pts) Give pseudocode or otherwise explain how you would use stochastic gradient descent to find $w^* = \arg\min_w J(w)$. You need to specify your approach to the step size, but you do not have to specify a stopping criterion, though you may if you like.

   Solution:

   - $t = 1$
   - Learning rate η =1
   - $w = 0$
   - Repeat until stopping criterion met:
     - randomly choose (x_i, y_i) from $\mathcal{D}$.
     - $w \leftarrow w - \eta\left(2\lambda w + \begin{cases} -2\left(1 - y_i w^T x_i\right) y_i x_i & y_i w^T x < 1 \\ 0 & \text{otherwise} \end{cases}\right)$

- $t \leftarrow t + 1$
- $\eta = 1/t$

4. (2 pts) Justify the claim that the output of SGD can be written in the form:

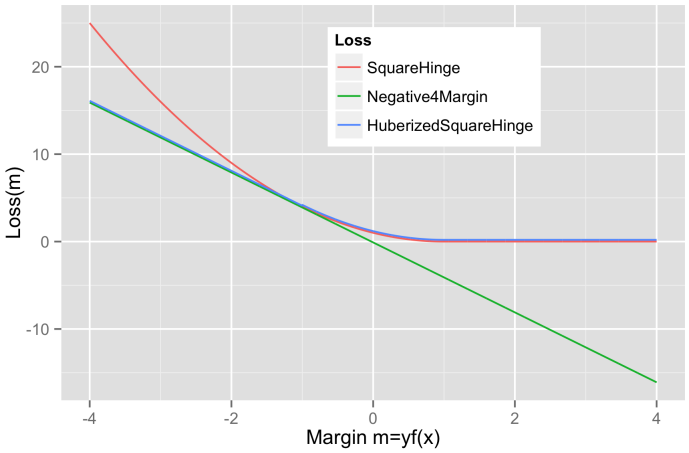
$$w = \sum_{i=1}^{n} \beta_i x_i.$$

5. (2 pts) In relation to the SGD algorithm, how would you characterize the x_i's that are support vectors?

6. (2 pts) The "Huberized" square hinge loss is a margin loss given by

$$\ell(m) = \begin{cases} -4m & m < -1 \\ \left[(1-m)_+\right]^2 & \text{otherwise.} \end{cases}$$

The plot below should help explain how this loss relates to the square hinge loss.



When might you prefer the Huberized square hinge loss to the square hinge loss?
Solution: Huberized square hinge loss is more robust to outliers.

Part II

2016 Test One

6 Step Directions in Optimization

1. [1] $J(w) : \mathbf{R}^d \to \mathbf{R}$ is a **convex**, **differentiable** objective function with minimizer w^*. Assume w_i is not a minimizer of $J(w)$. Let $g = \nabla J(w_i)$, 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: (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

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

and the dual optimization problem is

$$\sup_{\lambda \geq 0} \inf_{w} L(w, \lambda)$$

5. [1] $J(w)$ is not differentiable. Give an equivalent optimization problem that has a differentiable objective function. [Hint: You may want to introduce new variables as we did for the SVM.]
SOLUTION:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{such that} & \xi_i \geq 0 \ \forall i \\ & \xi_i \geq -y_i w^T x_i \leq 0 \ \forall i \\ & w^T w \geq 1 \end{array}$$

This would be sufficient. But we can also put it into standard form:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\ \text{such that} & -\xi_i \leq 0 \ \forall i \\ & -\xi_i - y_i w^T x_i \leq 0 \ \forall i \\ & 1 - w^T w \leq 0 \end{array}$$

6. [1] Is this a convex optimization problem? Why or why not?
SOLUTION: This is not a convex optimization problem. A convex optimization problem must have a convex feasible set. [Recall: The feasible set is the set that satisfies all the constraints.] The set of w satisfiying $\|w\|^2 \geq 1$ is not a convex set. You can also see this from the standard form in the previous problem. The function $1 - w^T w$ is concave, not convex.

7. [1] There's a good reason for the constraint $\|w\|^2 \geq 1$. What is the **unconstrained** minimizer of $J(w)$? Explain your answer. [Hint: This does not require calculation.]
SOLUTION: At $w = 0$, the objective function is 0. Since the objective function is always nonnegative, $w = 0$ is a solution to the unconstrained problem.

9 Regularization and the Bias Term

Suppose our input space is $\mathcal{X} = \mathbf{R}^2$ and our output space is $\mathcal{Y} = \mathbf{R}$. We have n labeled training points that we put together into a design matrix $X \in \mathbf{R}^{n \times 2}$. Let $y \in \mathbf{R}^n$ represent the vector of outputs. **This data stays fixed for all parts of this problem**. We'd like to find an affine function that fits this data. Let's create an "augmented design matrix" matrix $X_b \in \mathbf{R}^{n \times 3}$, whose first column is $b = (B, \ldots, B) \in \mathbf{R}^{n \times 1}$, for some $B > 0$. The next two columns are the original matrix X. Suppose X_b is full rank, and let

$$w^* = \arg\min_{w \in \mathbf{R}^3} \|X_b w - y\|^2.$$

Let's write the components of w^* as $w^* = (w_0, w_1, w_2)$.

1. [1] Given a new $x = (x_1, x_2)$, give an expression for the prediction on x corresponding to w^*. (Hint: This expression should be in terms of w_0, w_1, w_2, x_1, x_2 and B).
SOLUTION:
$$x \mapsto w_0 B + w_1 x_1 + w_2 x_2$$

2. [1] Let w^1 be the solution with $B = 1$ and let w^{100} be the solution with $B = 100$. Let w_0^1 and w_0^{100} refer to the first component of w^1 and w^{100} respectively. If $f_1(x)$ is the prediction corresponding to w^1 and $f_{100}(x)$ is the prediction corresponding to w^{100}, then for all x we have (choose one of the following): [Note: Assume $w_0^1 \neq 0$.]

   (a) $f_1(x) < f_{100}(x)$
   (b) $f_1(x) > f_{100}(x)$
   (c) $w_0^1 = w_0^{100}$and $f_1(x) = f_{100}(x)$
   (d) $\left|w_0^1\right| < \left|w_0^{100}\right|$ and $f_1(x) = f_{100}(x)$
   (e) $\left|w_0^1\right| > \left|w_0^{100}\right|$ and $f_1(x) = f_{100}(x)$

   SOLUTION: (e) The hypothesis space is the same regardless of the value of B, so without regularization, we will get $f_1(x) = f_{100}(x)$. Without regularization, we will have $w_0^{100} = \frac{1}{100} w_0^1$.

3. [1] Suppose for $B = 1$, the prediction function is $x \mapsto 5.3 + 2.0x_1 + 0.9x_2$. When we introduce an ℓ_2 regularization term, the resulting prediction function is $x \mapsto 1.99 + 2.10x_1 + 1.06x_2$. Now suppose we refit the model with ℓ_2 regularization but with $B = 100$. Which of the following prediction functions seems most likely to be the result (NOTE: these functions are coming from an iterative optimization algorithm and may have some optimization error):

   (a) $x \mapsto 1.99 + 2.10x_1 + 1.06x_2$
   (b) $x \mapsto 0.68 + 2.15x_1 + 1.13x_2$
   (c) $x \mapsto 5.49 + 1.96x_1 + 0.88x_2$
   SOLUTION: (c) There are 3 options: the bias term stays the same, gets bigger, or gets smaller. Without regularization, $w_0^{100} = \frac{1}{100} w_0^1$. So when $B = 100$, we expect the coefficient of B to be much smaller than when $B = 1$. Thus it decreases the amount of ℓ_2 penalty, which means we expect it to be closer to its unregularized value, which corresponds to a bias term of ~ 5.3. [*NOTE: This is not a great question, because it's not clear that you can reason your way to certainty that (c) is the correct answer. What we know for sure is that fitting with $B = 100$ is equivalent to fitting with $B = 1$ but with a reduction in the regularization on the bias term by a factor of B^2 – see solution to homework problem.*]

4. [1] Suppose $B = 1$. Let's introduce a new feature that is a duplicate of x_2, namely $x_3 = x_2$. Suppose we fit the model with ℓ_2 regularization. Which of the following prediction functions seems most likely to result? (NOTE: these functions are coming from an iterative optimization algorithm and may have some optimization error):

   (a) $x \mapsto 1.98 + 2.09x_1 + 0.20x_2 + 0.88x_3$
   (b) $x \mapsto 1.98 + 2.09x_1 + 0.54x_2 + 0.54x_3$

(c) Both seem equally likely
SOLUTION: (b). Both (a) and (b) give the same predictions, and thus have the same empirical loss. However, the ℓ_2 penalty for (b) is smaller, because $.54^2 + .54^2 < .2^2 + .88^2$. You don't need to calculate to know this: in general, if we minimize $a_1^2 + \cdots + a_d^2$ subject to $a_1 + \cdots + a_d = c$, the solution is $a_i = c/d$, for all $i = 1, \ldots, d$. You should know this result.

5. [1] Suppose $B = 1$. With **lasso** ℓ_1 regularization and the original feature set, the prediction function is $x \mapsto 4.47 + 2.05x_1 + 0.93x_2$. Let's introduce a duplicate feature $x_3 = x_2$. Suppose we refit with the same ℓ_1 regularization penalty. Which of the following prediction functions seems most likely to result? (NOTE: these functions are coming from an iterative optimization algorithm and may have some optimization error):

   (a) $x \mapsto 4.45 + 2.05x_1 + 0x_2 + 0.93x_3$

   (b) $x \mapsto 4.45 + 2.05x_1 + .93x_2 + 0x_3$

   (c) $x \mapsto 4.45 + 2.05x_1 + .46x_2 + .47x_3$

   (d) They are equally plausible results.
   SOLUTION: (d) Note that (a), (b), and (c) all give the same predictions, and all have the same ℓ_1 regularization term. Thus they are equivalent minimizers of the objective function. Lasso gives sparsity sometimes, but not always. When we have duplicate features, lasso doesn't care how the weight gets divided amont the duplicate features.

6. [1] Suppose $B = 1$. With **lasso** ℓ_1 regularization and the original feature set, the prediction function is $x \mapsto 4.47 + 2.05x_1 + 0.93x_2$. Let's introduce a new feature that is a **multiple** of x_2, namely $x_3 = 2x_2$. Suppose we refit with the same ℓ_1 regularization penalty. Which of the following prediction functions seems most likely to result? (NOTE: these functions are coming from an iterative optimization algorithm and may have some optimization error):

   (a) $x \mapsto 4.45 + 2.05x_1 + 0x_2 + 0.47x_3$

   (b) $x \mapsto 4.45 + 2.05x_1 + .94x_2 + 0x_3$

   (c) $x \mapsto 4.45 + 2.05x_1 + .71x_2 + .11x_3$

   (d) $x \mapsto 4.45 + 2.05x_1 + .34x_2 + .30x_3$

   (e) All seem equally likely
   SOLUTION: (a). All 4 give the same predictions, but (a) has the smallest ℓ_1 regularization penalty.

.1 Convexity

.1.1 Examples of Convex Functions (BV 3.1.5)

Functions mapping from $\mathbf{R}$:

- $x \mapsto e^{ax}$ is convex on $\mathbf{R}$ for all $a \in \mathbf{R}$
- $x \mapsto x^a$ is convex on $\mathbf{R}_{++}$ when $a \geq 1$ or $a \leq 0$ and concave for $0 \leq a \leq 1$
- $|x|^p$ for $p \geq 1$ is convex on $\mathbf{R}$
- $\log x$ is concave on $\mathbf{R}^{++}$
- $x \log x$ (either on $\mathbf{R}_{++}$ or on $\mathbf{R}_+$ if we define $0 \log 0 = 0$) is convex

Functions mapping from $\mathbf{R}^n$:

- Every norm on $\mathbf{R}^n$ is convex
- Max: $(x_1, \ldots, x_n) \mapsto \max\{x_1 \ldots, x_n\}$ is convex on $\mathbf{R}^n$

.1.2 Operations the preserve convexity (BV 3.2, p. 79)

Nonnegative weighted sums If $f_1, \ldots, f_m$ are convex and $w_1, \ldots, w_m \geq 0$, then $f = w_1 f_1 + \cdots + w_m f_m$ is convex. is convex in x (provided the integral exists).

Composition with an affine mapping A function $f : \mathbf{R}^n \to \mathbf{R}^m$ is an **affine function** (or **affine mapping)** if it is a sum of a linear function and a constaint. That is, if it has the form $f(x) = Ax + b$, where $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$.

Composition of a convex function with an affine function is convex. More precisely: suppose $f : \mathbf{R}^n \to \mathbf{R}$, $A \in \mathbf{R}^{n \times m}$ and $b \in \mathbf{R}^n$. Define $g : \mathbf{R}^m \to \mathbf{R}$ by $g(x) = f(Ax + b)$, with $\mathrm{dom} g = \{x \mid Ax + b \in \mathrm{dom} f\}$. Then if f is convex, then so is g; if f is concave, so is g. If f is **strictly** convex, and A has linearly independent columns, then g is also strictly convex.

Simple Composition Rules

- If g is convex then $\exp g(x)$ is convex.
- If g is convex and nonnegative and $p \geq 1$ then $g(x)^p$ is convex.
- If g is concave and positive then $\log g(x)$ is concave
- If g is concave and positive then $1/g(x)$ is convex.

Maximum of convex functions is convex (BV Section 3.2.3, p. 80) *Note: Below we use this to prove that the Lagrangian dual function is concave.*

If $f_1, \ldots, f_m : \mathbf{R}^n \to \mathbf{R}$ are convex, then their pointwise maximum

$$f(x) = \max\{f_1(x), \ldots, f_m(x)\}$$

is also convex with domain $\mathrm{dom} f = \mathrm{dom} f_1 \cap \cdots \cap \mathrm{dom} f_m$.

This result extends to the supremum over arbitrary sets of functions (including uncountably infinite sets).