6.7900: Exam 1, Fall 2024

Solutions

These are not the **only** acceptable answers. Some other answers also received credit.

Answer the questions in the spaces provided. Show your work neatly. **Try your best to put your answers in the boxes provided. If you absolutely have to write an answer elsewhere, mark very clearly where to find it.**

If a question seems vague or under-specified to you, make an assumption, write it down, and solve the problem given your assumption.

You may prepare and use both sides of one 8.5 inch x 11 inch sheet of paper upon which you may write/print anything you like. You may not use any electronic device or any other resource other than your two-sided sheet of paper.

Write your name on every page.

Try not to ask questions! But, if you feel you must, come up to the front.

Name: _____ MIT Email: _____

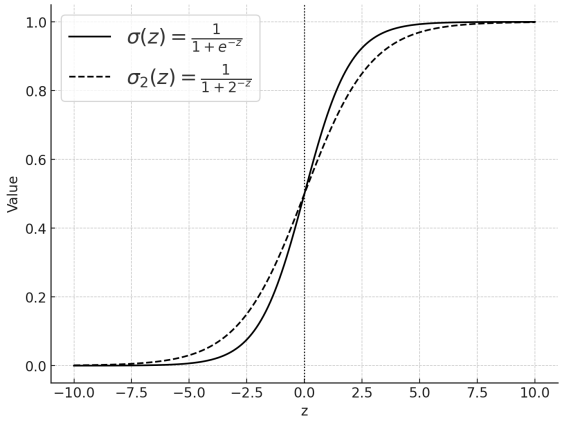
Question	Points	Score
1	24	
2	10	
3	25	
4	21	
5	20	
Total:	100	

Name: _____

1 Logistic Regression Progression

1. You have a training dataset $\mathcal{D}$ with each input $x^{(i)}$ consisting of d binary features $x_1^{(i)}, \ldots, x_d^{(i)}$, where all $x_j^{(i)} \in \{0, 1\}$ and a binary label $y^{(i)} \in \{0, 1\}$. You have been having trouble finding a good model, and you are getting low likelihood on the training set. Notation:

   - $[a, b, \ldots, z]$ concatenation
   - $\widehat{p^{(i)}}$: prediction for example i.
   - $\sigma(z) = 1/(1 + e^{-z})$, the sigmoid function
   - $\sigma_2(z) = 1/(1 + 2^{-z})$, the base-2 sigmoid function
   - Plots of $\sigma(z)$ and $\sigma_2(z)$ (for reference)



   (a) (6 points) Let's start by writing an expression for the training data likelihood, in terms of the training data and parameters $\theta \in \mathbb{R}^d$:

   Solution:

$$p(\mathcal{D}) = \prod_{i=1}^{n} \sigma(\theta^T x^{(i)})^{y^{(i)}} \cdot (1 - \sigma(\theta^T x^{(i)}))^{(1-y^{(i)})}$$

Now, here are some ideas for improving the training-set likelihood. Let's assume that we are always finding the optimum of the likelihood given the features we are using. For each of them, specify whether: 1) it generally increases training-set likelihood; 2) it will not make a difference; or 3) it generally decreases training-set likelihood. **Explain each answer in one or two sentences or formulas.**

(b) (3 points) For each data-point, augment the dimensions by adding the *complement* of each feature before trying to fit the dataset; i.e. $[x_1^{(i)}, \ldots, x_d^{(i)}] \rightarrow [x_1^{(i)}, \ldots, x_d^{(i)}, 1 - x_1^{(i)}, \ldots, 1 - x_d^{(i)}]$

training-set likelihood generally: ◯ increases √ **doesn't change** ◯ decreases

Solution: Any solution to this new logistic regression can be parametrized as $\widehat{y^i} = \sigma\left(\sum_{j=1}^{f} \alpha_j x_j^i + \sum_{j=1}^{f} \beta_j (1 - x_j^i) + \gamma\right)$, then using the distributive property:

$$\widehat{y^i} = \sigma\left(\sum_{j=1}^{f} \alpha_j x_j^i + \sum_{j=1}^{f} \beta_j (-x_j^i) + \sum_{j=1}^{f} \beta_j \cdot 1 + \gamma\right)$$

$$\widehat{y^i} = \sigma\left(\sum_{j=1}^{f} (\alpha_j - \beta_j) x_j^i + \left(\sum_{j=1}^{f} \beta_j + \gamma\right)\right)$$

which has the form of the original logistic regression.

It was enough to say that because the new features are a linear function of the original ones, they do not add expressive power.

Note: Full credit was also given for good reasoning that this feature transformation effectively adds a bias term if it didn't previously exist in the model, generally increasing likelihood.

(c) (3 points) Use two sigmoids instead of one; i.e., instead of parametrizing the solution as $\widehat{p^{(i)}} = \sigma(\sum \theta_j x_j^{(i)})$, parametrize it as: $\widehat{p^{(i)}} = \sigma(\sigma(\sum \theta_j x_j^{(i)}))$.

training set likelihood generally: ◯ increases ◯ doesn't change √ **decreases**

Solution: The output of $\sigma(z)$ is between 0 and 1 for all z. Therefore the output of $\sigma(\sigma(z))$ is between $\sigma(0) = 0.5$ and $\sigma(1) \approx 0.73$, which severely limits its range and in particular cannot predict $y^i = \texttt{False}$.

(d) (3 points) Fit a regular logistic regression using the base e sigmoid, σ, and a second logistic regression using a base-2 sigmoid, σ_2, function and return the most confident result (the probability estimate that is farthest from 0.5).

training set likelihood generally: ◯ increases √ **doesn't change** ◯ decreases

Solution: Assuming that we denote the model logits $\sum_{j=1}^{f} \alpha_j x_j^i + \gamma$, we can derive the solution using σ_2 and σ:

$$\sigma_2(\sum_{j=1}^{f} \alpha_j x_j^i + \gamma) = \frac{1}{1 + 2^{-(\sum \alpha_j x_j^i + \gamma)}} = \frac{1}{1 + e^{-\ln 2(\sum x_j^i + \gamma)}} = \sigma(\sum_{j=1}^{f} \alpha_j x_j^i \ln 2 + \gamma \ln 2)$$

The factor $\ln 2$ can be incorporated into the learned weights. Therefore, σ_2 logistic regression and σ logistic regression are equally expressive and training them will yield equal predictions for all datapoints. Therefore ensembling them in the proposed manner will not change results.

(e) (3 points) Before fitting the dataset, for every data-point i, make an independent coin flip and append the result of that flip as a feature; i.e., $[x_1^{(i)}, \ldots, x_d^{(i)}] \to [x_1^{(i)}, \ldots, x_d^{(i)}, \text{coin}(i)]$.

training set likelihood generally √ **increases** ◯ doesn't change ◯ decreases

Solution: First, it is clear that this function class is at least as expressive because we can always set the coefficient multiplying $\texttt{coin}(i)$ to 0. Then, with high probability, the random feature will have non-0 correlation with the labels and will thus contain some information about them that can be exploited to reduce underfitting. More concretely, for n datapoints the probability of this happening is $\frac{\binom{n}{n/2}}{2^n} \to 0$ and 0 for n odd, this level of detail was not needed for a satisfactory answer.) Another way of seeing it is that $\texttt{coin}(i)$ is a function that is not a linear combination of the current feature set and will thus increase the capacity of our classifier.

Note that this would be a bad idea because it would lead to poor generalization, but it would still help with underfitting.

Name: _____

(f) (3 points) Fit a regular logistic regression and obtain a prediction $\widehat{p^{(i)}}$ for each element $x^{(i)}$. Append $\widehat{p^{(i)}}$ as a feature ($[x_1^{(i)}, \ldots, x_d^{(i)}, \widehat{p^{(i)}}]$) and fit another logistic regression to the new dataset.

training set likelihood generally: $\surd$ **increases** ◯ doesn't change ◯ decreases

Solution: $\widehat{y_i}$ is a non-linear function combination of the features and therefore increases the capacity of the logistic regresssion. Moreover, $\widehat{y_i}$ contains information on the labels through the parameters trained during the first logistic regression, which makes it a very informative feature.

(g) (3 points) For each pair of features add their product as another feature before trying to fit the dataset; i.e. $[x_1^{(i)}, \ldots, x_d^{(i)}] \rightarrow [x_1^{(i)}, \ldots, x_d^{(i)}, x_1^{(i)}x_2^{(i)}, x_1^{(i)}x_3^{(i)}, \ldots, x_{d-1}^{(i)}x_d^{(i)}]$.

training set likelihood generally: $\surd$ **increases** ◯ doesn't change ◯ decreases

Solution: We can express anything that we could express before by setting all coefficients for second-order terms to 0, but we now introduce nonlinear interaction features that could better fit the training data.

Name: _____

2 Ridges

2. Recall the ridge regression objective

$$J(\theta) = \sum_{i=1}^{n} (\theta^T x^{(i)} - y^{(i)})^2 + \lambda \|\theta\|^2 \ .$$

We saw in the homework that ridge regression can be understood as a maximum *a posteriori* probability (MAP) estimator in the case where the prior on θ is a Gaussian with mean vector $\mathbf{0}$ and covariance matrix $\alpha\mathbf{I}$ for some scalar value α.

Recall that the posterior probability distribution on θ given data $\mathcal{D}$, $p(\theta \mid \mathcal{D})$, is proportional to $p(\mathcal{D} \mid \theta)p(\theta)$.

(a) (6 points) How would we rewrite the ridge regression objective if our prior had a mean vector w instead of $\mathbf{0}$?

Solution:

$$J(\theta) = \sum_{i=1}^{n} (\theta^T x^{(i)} - y^{(i)})^2 + \lambda \|\theta - w\|^2 \ .$$

(b) (4 points) Suppose we were to increase α. What change, if any, would we need to make to our objective, so that its minimizer is the MAP of the resulting distribution? Choose all that apply and **give an informal explanation**.

√ **Decrease** λ ◯ Increase λ ◯ Decrease $\|w\|$ ◯ Increase $\|w\|$

Solution: Recall from HW2 Q5 that λ is inversely proportional to α, so increasing α would decrease the regularization constant. Intuitively, if we are less certain about our prior of the weights, we should penalize less for deviating from the prior.

Name: _____

3 Discrete Bayesian Logistic Regression

3. We are performing logistic regression, in the sense that we are interested in fitting a probability distribution $P(y \mid x)$ of the form

$$p(y = 1 \mid x, \theta) = \sigma(\theta^T x) \ ,$$

where $y \in \{0, 1\}$ and $x \in R^d$ is an input vector, to a data set of iid (x, y) pairs.

Often, in logistic regression, we have no prior on θ, and we simply try to a maximum-likelihood estimate for θ. In this problem, by contrast, $d = 2$ and for some reason we are sure that there are only 4 possible values for θ: $\theta_1 = (1, 1), \theta_2 = (1, -1), \theta_3 = (-1, 1), \theta_4 = (-1, -1)$. We start with some discrete prior distribution over θ, $p(\theta) = (p_1, p_2, p_3, p_4)$.

(a) (6 points) Write a formula for the prior *marginal probability* $p(y = 1 \mid x)$ in this model, just in terms of σ, the θ_i and prior p_i values for $i \in \{1, \ldots, 4\}$, and x.

Solution:

$$p(y = 1|x) = \sum_{i=1}^{4} p(y = 1|x, \theta = \theta_i) p(\theta = \theta_i)$$
$$= \sum_{i=1}^{4} p_i \sigma(\theta_i^T x)$$

(b) (6 points) Write a general formula for the posterior on parameter values after obtaining one observation (x, y). You can ignore any constants of proportionality and just specify $p(\theta = \theta_i \mid (x, y)) \propto$

Solution: $P(\theta = \theta_k \mid (x, y)) \propto P(y|x, \theta_k) P(\theta = \theta_k)$.
So we can write

$$P(\theta = \theta_k \mid (x, y)) \propto \sigma(\theta_k^T x)^y (1 - \sigma(\theta_k^T x))^{1-y} p(\theta = \theta_k)$$

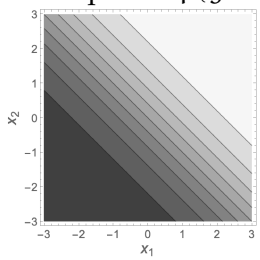
(c) (6 points) Now, assume our prior on these hypotheses is uniform, and imagine you get an input of $(0, 1)$ and it is labeled positive. After you perform a Bayesian update based on this observation, for each of the hypotheses, what happens:

i. $\theta_1 = (1, 1)$ becomes $\surd$ **more likely** ◯ less likely ◯ unchanged
ii. $\theta_2 = (1, -1)$ becomes ◯ more likely $\surd$ **less likely** ◯ unchanged
iii. $\theta_3 = (-1, 1)$ becomes $\surd$ **more likely** ◯ less likely ◯ unchanged
iv. $\theta_4 = (-1, -1)$ becomes ◯ more likely $\surd$ **less likely** ◯ unchanged

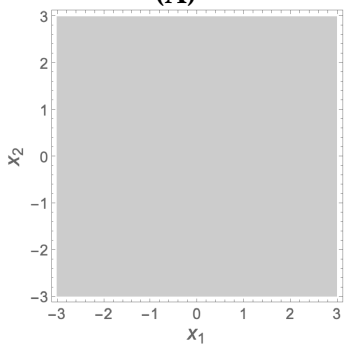
(d) (5 points) Each of the plots (A)–(E) shows, for each point $x = (x_1, x_2)$, a value $p(y = 1 \mid x)$, where more likely areas are brighter. For example, the upper right corner is more likely

Name: _____

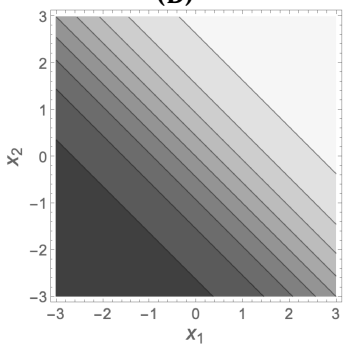
in the plot of $p(y = 1 \mid x, \theta = (1, 1))$ below



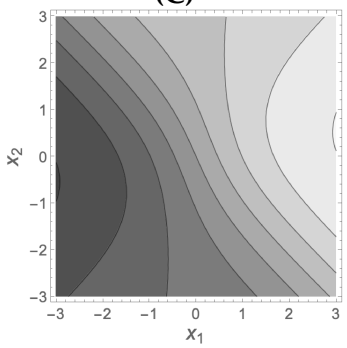
(A)



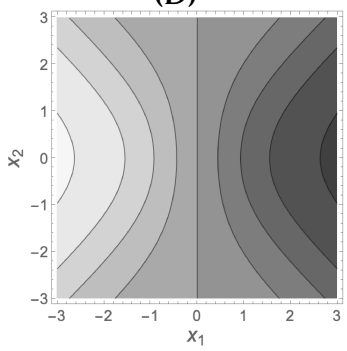
(B)



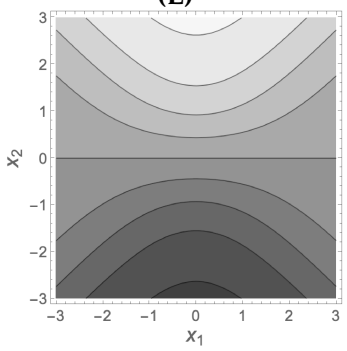
(C)



(D)



(E)



Match each of these prior or posterior predictive distributions, given data D, to the appropriate figure (again assuming the prior on parameters is uniform).

i. $p(y \mid x, D = \{((1,0),0)\})$
◯ A ◯ B ◯ C √ **D** ◯ E

ii. $p(y \mid x, D = \{\ \})$
√ **A** ◯ B ◯ C ◯ D ◯ E

iii. $p(y \mid x, D = \{((1,0.5),1)\})$
◯ A ◯ B √ **C** ◯ D ◯ E

iv. $p(y \mid x, D = \{((1,0),1),((0,1),1)\})$
◯ A √ **B** ◯ C ◯ D ◯ E

v. $p(y \mid x, D = \{((0,1),1)\})$
◯ A ◯ B ◯ C ◯ D √ **E**

(e) (2 points) Briefly explain your answers.

Solution: Explanation:
(i). d is the only option symmetric around $x_2 = 0$ and higher for x_1 negative
(ii). a is the only option where all four corners have the same probability
(iii). The upper right corner has the most probability in both b and c. but in c, we don't have symmetry around $x_1 = x_2$
(iv). see answer for iii. here we do expect symmetry around $x_1 = x_2$
(v). e is the only option symmetric around $x_1 = 0$ and higher for x_2 positive

Name: _____

4 Fighting last week's fire

4. In this question, we will revisit the problem posed in mini-project 1. Recall that we have data that associates several features of the context in which the fire is occurring (x) with its eventual total burned area (assuming no intervention), y. Our goal was to predict the best fire-fighting response, to minimize an overall economic cost, including the cost of the response and the cost of the burned acreage. We defined $\mathit{risk}(y, r)$ to be the *expected* economic cost of sending response r to a fire with (untreated) burn acreage y.

   We discussed three different framings of the problem:

   1. **Basic regression**: try to find parameters θ of a regression model that predicts $y = h_\theta(x)$ with low squared loss. Let f^* be the linear regression model that minimizes true expected squared error.
   2. **Response classification**: assign each training input $x^{(i)}$ to a discrete response in some finite set R, where $r^{(i)} = \text{argmin}_{r \in R} \mathit{risk}(y^{(i)}, r)$, and try to find parameters θ of a multi-class classifier to predict r, minimizing cross-entropy loss or 0-1 loss with respect to these assigned classes. Let c^* be the classifier that minimizes true expected 0-1 loss.
   3. **Cost-based classification**: try to find parameters θ of a classification model that predicts a response $r \in R$ (where R is a discrete set of responses) that minimizes $L_E(x, r)$, which is the expected cost of sending response r to a fire with properties x. Let r^* be the classifier that minimizes the true value of this cost-based loss.

   (a) (5 points) Give a definition for r^* in terms of $\mathit{risk}(y, r)$ and the true $p(y \mid x)$.

   Solution:

   $$r^*(x) = \text{argmin}_r \int_y p(y \mid x) \mathit{risk}(y, r) dy$$

(b) (3 points) Jan thinks that $r^*(x)$ and $\text{argmin}_{r \in R} risk(f^*(x), r)$ are not equal. For some particular x, let $p(y \mid x) = unif(0, 2)$ and $f^*(x) = 1$. Assume there are only two possible responses, and that $risk(y, r_1) = 1$. Describe, in math or words, a function $risk(y, r_2)$ for which $r^*(x)$ is not equal to $\text{argmin}_r risk(f^*(x), r)$ and where r_1 is a better response.

Solution: First of all, this question should have just said that we should construct a risk function such that:

$$risk(f^*(x), r_1) > risk(f^*(x), r_2)$$
$$r^*(x) = r_1$$

So that picking according to the best regression result would pick r_1 but picking according to the integrated risk would pick r_2.

This will happen when the risk for r_2 is is less than 1 (which is the risk for r_1) at $y = 1$, but where the integrated risk is higher than 2.

Here is an example of such a function.

$$risk(y, r_2) = y^4 - .1$$

But really it could be anything that goes up once it's away from y = 1.

(c) Lauri decides to apply Bayesian linear regression to the problem and finds that, for the next fire that crops up, the posterior predictive distribution has very low variance. They enthusiastically go and tell everybody that they are highly certain about the regression result.

i. (2 points) Describe a situation (property of the data) that would cause Lauri's certainty to be mis-placed, even if this new fire is drawn from the same distribution as the previous ones?

Solution: If Lauri's assumptions for Bayesian linear regression are correct, then they should be right in being certain if their posterior predictive has low variance. Things go wrong when their assumptions are broken, such as the true model not being linear.

ii. (2 points) What additional tests on existing data could Lauri have run that would have prevented such over-optimism?

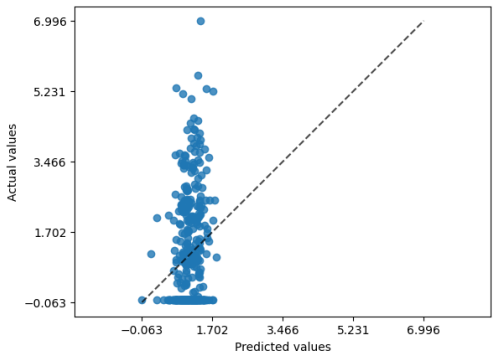
Solution: Look at data likelihood (of training and or a validation set), or visualize the data

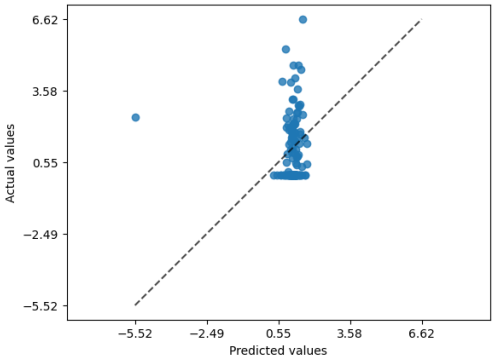
Name: _____

(d) (4 points) Here are training and evaluation plots of the residuals of four different regressors, on real firefighting data (in the project we looked at synthetic data). The y values are computed as $\log(1+a)$ where a was the area burned in the original dataset. By far the most common value of a was 0, and so the most common value of $\log(1+a)$ is also 0. We are also including some evaluation results for each one.

Regressor 1: train (left plot), evaluation (right plot)

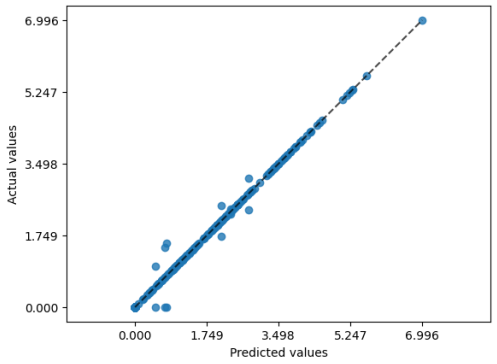
Notice that the plots have different scales

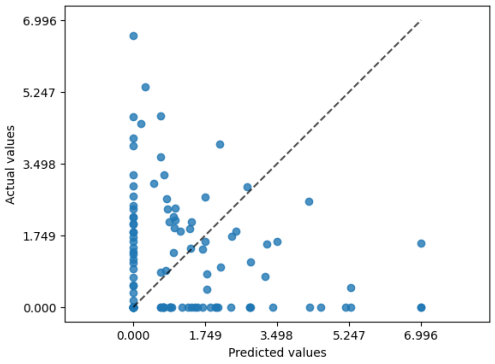





```
Training RMSE 1.37                Evaluation RMSE 1.54
Training MAD 1.05                 Evaluation MAD 1.09
Training empirical risk 184415    Evaluation empirical risk 207163
```

Regressor 2: train (left plot), evaluation (right plot)

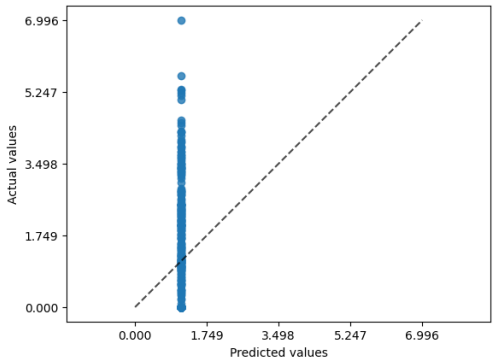


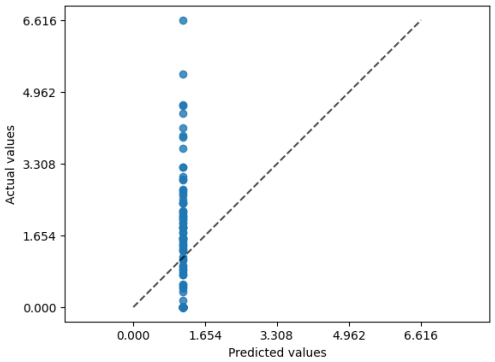



```
Training RMSE 0.09                Evaluation RMSE 2.18
Training MAD 0.0                  Evaluation MAD 1.03
Training empirical risk 57857     Evaluation empirical risk 237487
```

Name: _____

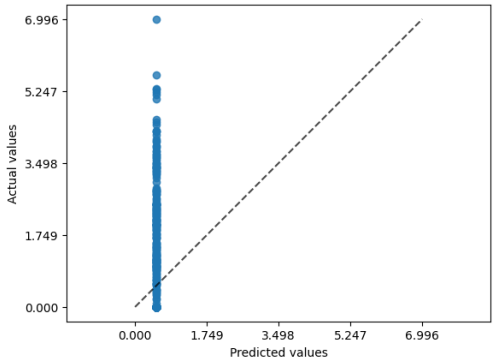
Regressor 3: train (left plot), evaluation (right plot)

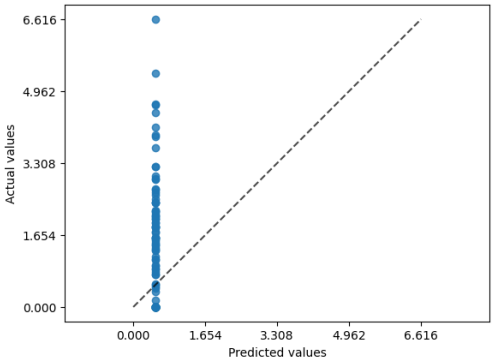





```
Training RMSE 1.40                Evaluation RMSE 1.38
Training MAD 1.13                 Evaluation MAD 1.13
Training empirical risk 192331    Evaluation empirical risk 214310
```

Regressor 4: train (left plot), evaluation (right plot)






```
Training RMSE 1.52                Evaluation RMSE 1.48
Training MAD 0.51                 Evaluation MAD 0.51
Training empirical risk 192331    Evaluation empirical risk 214310
```

Match each of the four regression methods to the resulting data from above.

i. Predict the mean y value
   ◯ Regressor 1 ◯ Regressor 2 √ **Regressor3** ◯ Regressor 4

ii. Predict the median y value
   ◯ Regressor 1 ◯ Regressor 2 ◯ Regressor3 √ **Regressor 4**

iii. Ordinary linear regression
   √ **Regressor 1** ◯ Regressor 2 ◯ Regressor3 ◯ Regressor 4

iv. Decision-tree regression with a large tree
   ◯ Regressor 1 √ **Regressor 2** ◯ Regressor3 ◯ Regressor 4

(e) (2 points) Something looks suspicious here: why are the empirical risks for regressor 3 and regressor 4 the same?

> **Solution:** They are both predicting a very small fire size, which generates a very small response, to all cases.

(f) (3 points) Which regressor would you use to predict responses to upcoming fires? Give a brief explanation. √ **Regressor 1** ◯ Regressor 2 ◯ Regressor3 ◯ Regressor 4

> **Solution:** It has the lowest empirical risk on the validation data.
> But fine if they say they're all terrible!

Name: _____

5 Variations on a theme

5. We are observing an important scientific phenomenon, but find that our instrument is noisier during the day than during the evening. In particular, we believe that our ith observation $y^{(i)}$ is a deterministic scalar function of some feature vector $x^{(i)}$ plus some zero-mean Gaussian noise, ϵ_i, that is independent across observations i: $y^{(i)} = f(x^{(i)}) + \epsilon_i$. Assume that the variance of ϵ_i during the day is σ^2_{day} and the variance of ϵ_i at night is σ^2_{night}, and both variances are known. Also, the first feature (that is, the first component of the feature vector, $x_1^{(i)}$) is strictly greater than 0 in the day time and less than or equal to 0 at night.

   Given a data set $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, we would like to fit a linear regression model, $f(x) = \theta^T x$. For simplicity in expressing your answer, assume datapoints $1, \ldots, m$ have $x_1 > 0$ and datapoints $m+1, \ldots, n$ have $x_1 \leqslant 0$.

   (a) (7 points) Derive a formula for the maximum likelihood estimate of θ, just expressed as an *argmax*. Recall the scalar Gaussian density with mean μ and σ^2:

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

Solution: *Key points:* write θ in terms of argmin or argmax, and include σ in the expression. Correctly express the log likelihood (or likelihood).

$$\log p(\mathcal{D}|\theta) = -\tfrac{1}{2}\sum_{i=1}^n \left(\log(2\pi\sigma_i^2) + \frac{((y^{(i)} - \theta^T x^{(i)})^2}{\sigma_i^2}\right)$$

$$\hat{\theta} = \operatorname*{argmin}_\theta \sum_{i=1}^n \frac{(y^{(i)} - \theta^T x^{(i)})^2}{\sigma_i^2}$$

or

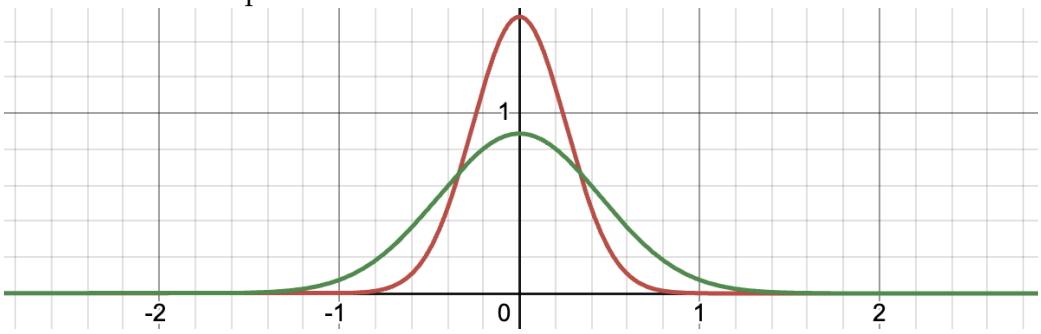
$$\hat{\theta} = \operatorname*{argmax}_\theta -\sum_{i=1}^n \frac{(y^{(i)} - \theta^T x^{(i)})^2}{\sigma_i^2}$$

(b) (6 points) For each part, mark True/False and briefly explain your answer.

i. ◯ True √ **False** : Suppose I have two observations such that $y^{(1)} - \theta^T x^{(1)} = y^{(2)} - \theta^T x^{(2)}$, but observation 1 is taken during the day and observation 2 is taken at night; and suppose $\sigma^2_{\text{day}} > \sigma^2_{\text{night}}$, then the likelihood of observation 1 is higher than the likelihood of observation 2, given θ.

Solution: The correct answer is **False** because we are comparing likelihoods of two different distributions at an arbitrary value, so which likelihood is bigger actually depends on which point we compare them at.
A common misconception was "higher noise means lower likelihood since the distribution is flatter" which is misplaced intuition that only works for values in the vicinity of zero.

*The **main thing to internalize in this solution** is that normal distributions with higher variance have relatively **lower likelihood near zero** but **higher likelihood far from zero*** as seen in the picture below.



Side note: One way to see that one likelihood is not guaranteed to be larger than the other is by 1) noticing that both the pdfs of values during the day and values during the night integrate to 1 and 2) realizing that if one was always smaler than the other, then that would bound the area under one of them to be smaller than 1, contradiction. e

ii. √ **True** ◯ False : Under the assumption that $\sigma_{\text{day}} = 2\sigma_{\text{night}}$, adding three extra copies of every nighttime reading and using the ordinary least squares solution will have the same effect as directly optimizing the objective from part (a)

Solution: All we are looking to do in this problem is imitate a weighted OLS objective with an unweighted OLS objective with modified data.
Looking at our LL objective derived in a), each datapoint's contribution to the loss is weighed by $\frac{1}{\sigma^2_{\text{datapoint}}}$, so points with halved std (night observations) have 4 times the influence.
In order for us to weight the nighttime observations $4\times$ without explicitly adding weights, we just $4\times$ each nighttime observation in our training data.

(c) (7 points) Lane comes from the lab across the hall with an even more difficult problem. They think that their output measurements would be exactly linear as a function of the features if they could measure the features. But in their lab, there is error in the process of sensing the features x. For simplicity, we'll think just about the case where x is one-dimensional. Then, an appropriate generative model is

$$y^{(i)} = \theta(x^{(i)} + \epsilon_i)$$

where the ϵ_i are i.i.d. across i with $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Suppose σ^2 is known.

It might be helpful to know that for a random variable X with mean μ and standard deviation σ, the random variable $Y = cX$ (for scalar constant c) has mean $c\mu$ and standard deviation $c\sigma$.

Derive a formula for the maximum likelihood estimator for θ, expressed as an *argmin* of a formula involving σ, θ, and the data values.

Solution: We have

$$p(y \mid x) \sim \mathcal{N}(\theta x, \theta^2\sigma^2)$$

Given a data set, our objective is to maximize the likelihood, or equivalently minimize the negative log likelihood

$$\begin{aligned}
\hat{\theta} &= \text{argmin} - \sum_{i=1}^{n} \log p(y^{(i)} \mid x^{(i)}) \\
&= \text{argmin}_\theta - \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma\theta} \exp\left(-\frac{1}{2\sigma^2\theta^2}(\theta x^{(i)} - y^{(i)})^2\right) \\
&= \text{argmin}_\theta \sum_{i=1}^{n} \log(\theta) + \frac{1}{2\sigma^2\theta^2}(\theta x^{(i)} - y^{(i)})^2 \\
&= \text{argmin}_\theta \left(n \log(\theta) + \frac{1}{2\sigma^2\theta^2} \sum_{i=1}^{n} (\theta x^{(i)} - y^{(i)})^2\right)
\end{aligned}$$

Name: _____

Work space

Name: _____

Work space