CS 6140: Machine Learning — Fall 2021— Paul Hand

Midterm 2 Study Guide and Practice Problems
Due: Never.

Names: [Put Your Name(s) Here]

This document contains practice problems for Midterm 1. The midterm will only have 5 problems. The midterm will cover material up through and including the bias-variance tradeoff, but not including ridge regression. Skills that may be helpful for successful performance on the midterm include:

If opt prob is Quadratic & convex, set $\nabla f = 0$

1. Write down the optimization problem corresponding to MAP estimation under a Bayesian Prior.
2. Solve the optimization problem corresponding to MAP estimation, in cases where this is possible.
3. Be able to state and prove the condition for convergence of gradient descent with a constant step size in the case of a quadratic function.
4. Write down an analytical expression for the solution to least squares problems with and without quadratic regularization terms.
5. Explain the behavior of the solutions to ridge regression for various values of regularization parameter λ, including relating the problem to overfitting, underfitting, bias, complexity, and convexity.
6. Explain the behavior of the solutions to k-nearest neighbors for regression and classification for various values of the parameter k, including relating the problem to overfitting, underfitting, and bias.
7. Compute the predictions for a k-nearest neighbor algorithm given a provided data set.
8. Implement cross validation for a provided data set and model.
9. Identify if a quadratic function is convex.

MLE $\underset{\theta}{\operatorname{argmax}}\ P(X|\theta)$

MAP $\underset{\theta}{\operatorname{argmax}}\ P(\theta|X)$

$$\underset{\theta}{\operatorname{argmax}}\ \log P(\theta|X) =$$

$$\underbrace{\boxed{\underset{\theta}{\operatorname{argmax}}\ \log P(X|\theta) + \log P(\theta)}}_{\text{MAP problem}} \cancel{-\log P(X)}$$

3)

3. Be able to state and prove the condition for convergence of gradient descent with a constant step size in the case of a quadratic function.

Quad. function: $f(x) = x^t Q x$ $\quad \nabla f(x) = 2Qx$

For GD to converge Q should have non neg. E-values

$$x^{(n+1)} = x^{(n)} - \alpha \nabla f(x^{(n)})$$

$$= x^{(n)} - \alpha 2Q x^{(n)}$$

$$= (I - 2\alpha Q)\, x^{(n)}$$

So $x^{(n)} = (I - 2\alpha Q)^n x^{(0)}$

$X^{(n)} \to 0$ if all eigenvalues of

$I - 2\alpha Q$ are s.t $-1 < \lambda_i(I - 2\alpha Q) < 1$

If $Q = U \Lambda U^t$ (eigenvalue decomp of Q)

$$I - 2\alpha Q = U(I - 2\alpha\Lambda)U^t$$

$$(I - 2\alpha Q)^n = U(I - 2\alpha\Lambda)^n U^t$$

4. Write down an analytical expression for the solution to least squares problems with and without quadratic regularization terms.

$$\arg\min_{\theta} \frac{1}{2}\|X\theta - y\|^2 + \lambda\|\theta\|^2$$

$$+ \lambda \sum_i c_i \theta_i^2$$

$$+ \theta^t Q \theta$$

$$\nabla_\theta f = \frac{1}{2} \cdot 2\, X^t(X\theta - y) + \cdots\cdots$$

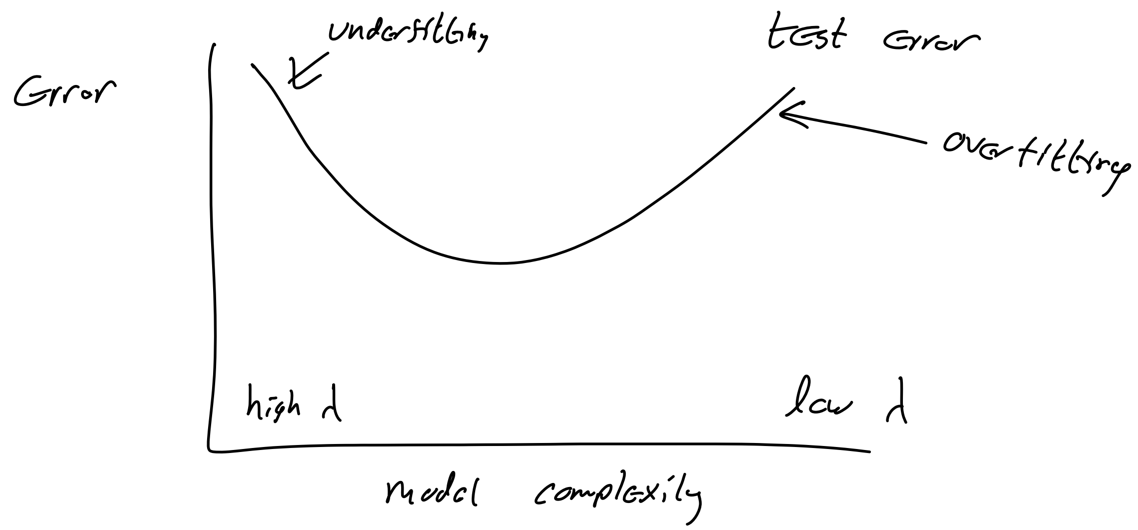
$2\lambda\theta$

or

$\vdots$

$2Q\theta$

5. Explain the behavior of the solutions to ridge regression for various values of regularization parameter λ, including relating the problem to overfitting, underfitting, bias, complexity, and convexity.



$$\arg\min_{\theta} \|y - X\theta\|^2 + \lambda \|\theta\|^2$$

Convexity

If f is (twice) differentiable

$f: \mathbb{R}^d \to \mathbb{R}$ is convex

if and only if

$H f(x)$ is positive semidefinite for all x

H is PSD if and only if $z^t H z \geq 0$ for all $z \in \mathbb{R}^d$

6. Explain the behavior of the solutions to k-nearest neighbors for regression and classification for various values of the parameter k, including relating the problem to overfitting, underfitting, and bias.

Question 1. *Maximum A Posteriori Estimation*

Suppose $y_i \sim \mathcal{N}(\mu, 1)$ for $i = 1 \ldots n$. Suppose μ has a Bayesian prior given by a Uniform$[-1, 1]$ distribution. Given the following data, find the MAP estimate of μ.

i	y_i
1	-1.5
2	-1.1
3	-0.5

Let $Y = \begin{pmatrix} -1.5 \\ -1.1 \\ -0.5 \end{pmatrix}$ $\qquad P(y_i|\mu) = \frac{1}{\sqrt{2\pi}} e^{-(y_i-\mu)^2/2}$

Response:

$$\log P(\mu|Y) = \log P(Y|\mu) + \log P(\mu) - \log P(Y)$$

$$= \sum_{i=1}^{3} \log P(y_i|\mu) + \log P(\mu) - \log P(Y)$$

$$= \sum_{i=1}^{3} \left[-\frac{(y_i-\mu)^2}{2} - \log\sqrt{2\pi} \right] + \log P(\mu) - \log P(Y).$$

Note: $P(\mu) = \begin{cases} 1/2 & \text{if} \quad -1 \leq \mu \leq 1 \\ 0 & \text{if} \quad \text{otherwise} \end{cases}$

$$\log P(\mu) = \begin{cases} -\log 2 & \text{if} \quad -1 \leq \mu \leq 1 \\ -\infty & \text{if} \quad \text{otherwise} \end{cases}$$

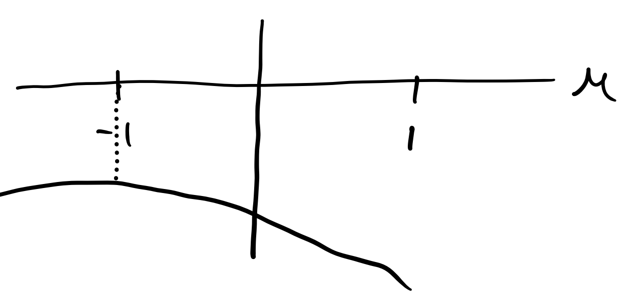
MAP Estimate is given by

$$\underset{\mu}{\operatorname{argmax}} \sum_{i=1}^{3} \left(-\frac{(y_i-\mu)^2}{2} - \log\sqrt{2\pi} \right) + \log P(\mu) - \log P(Y)$$

$$= \underset{-\infty < \mu < \infty}{\operatorname{argmax}} \sum_{i=1}^{3} -\frac{(y_i-\mu)^2}{2} + \log P(\mu)$$

$$= \underset{-1 \leq \mu \leq 1}{\operatorname{argmax}} \sum_{i=1}^{3} -\frac{(y_i-\mu)^2}{2}$$

Plotting this function



We see the max is at

$\boxed{\mu = -1}$

Question 2. *Maximum A Posteriori Estimation and Logistic Regression*

Consider the task of building a binary classifier. You have a training dataset $\{(x_i, y_i)\}_{i=1...n}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. Consider the statistical model where $P(y = 1 \mid x) = \sigma(\theta^t x)$, where σ is the logistic function. Write down the optimization problem that would be solved to perform MAP estimation of θ provided that θ has a prior distribution where each component θ_i is independent and normally distributed with mean 0 and variance σ^2.

Response:

Let $S = \{(x_i, y_i)\}_{i=1 \sim n}$

$$\log P(\theta|S) = \log P(S|\theta) + \log P(\theta) - \log P(S)$$

$$= \sum_{i=1}^{n} \log P((x_i, y_i)|\theta) + \log P(\theta) - \log P(S)$$

$$= \sum_{i=1}^{n} [\log P(y_i|x_i,\theta) + \log P(x_i|\theta)] + \log P(\theta) - \log P(S)$$

as x_i does not depend on θ

$$= \sum_{i=1}^{n} [\log P(y_i|x_i,\theta) + \log P(x_i)] + \log P(\theta) - \log P(S)$$

$$= \sum_{i=1}^{n} \log P(y_i|x_i,\theta) + \log P(\theta) + (\text{terms constant in } \theta)$$

$$P(\theta_i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\theta_i^2/2\sigma^2}$$

$$= \sum_{i=1}^{n} \left(1_{y_i=1} \cdot \log \sigma(\theta^t x_i) + 1_{y_i=0} \log(1-\sigma(\theta^t x_i))\right) + \log P(\theta) + (\text{constant in } \theta)$$

Note $P(\theta) = \frac{1}{\sqrt{2\pi}^d \sigma^d} e^{-\sum_{j=1}^{d} \theta_i^2/2\sigma^2}$

$$\Rightarrow \log P(\theta) = -\frac{\|\theta\|^2}{2\sigma^2} + \text{terms constant in } \theta$$

So MAP estimate is given by

$$\boxed{\arg\max_{\theta} \sum_{i=1}^{n} 1_{y_i=1} \cdot \log \sigma(\theta^t x_i) + 1_{y_i=0} \log(1-\sigma(\theta^t x_i)) - \frac{\|\theta\|^2}{2\sigma^2}}$$

Question 3.

(a) Show that for any matrix $X \in \mathbb{R}^{n\times d}$, XX^t and X^tX are positive semidefinite.

Response: We show $z^t X X^t z \geq 0$ for all $z \in \mathbb{R}^n$

observe $z^t X X^t z = \|X^t z\|^2 \geq 0.$

Similarly, for any $z \in \mathbb{R}^d$

$$z^t X^t X z = \|Xz\|^2 \geq 0$$

(b) Show that $\lambda_{\max}(X^tX) = \sigma^2_{\max}(X)$, where $\lambda_{\max}$ is the largest eigenvalue of X and $\sigma_{\max}$ is the largest singular value of X. Hint: Use a singular value decomposition of X in order to get an eigenvalue decomposition of X^tX.

Response:

Let $X = U\Sigma V^t$ be an SVD of $X \in \mathbb{R}^{n\times d}$

where U has orthonormal columns
V has orthonormal columns
Σ is diagonal w/ nonnegative entries.

We have
$$X^tX = V\Sigma^t U^t U \Sigma V^t$$
$$= V\Sigma^t I \Sigma V^t$$
$$= V\Sigma^2 V^t,$$

which is an eigenvalue decomposition.
So the eigenvalues of X^tX are given by the diagonal entries of Σ^2.

As X^tX is positive semidefinite, its eigenvalues are nonnegative. So $\lambda_{\max}(X^tX) = \sigma^2_{\max}(X)$.

Question 4. *Ridge Regression*

Let $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$, $\lambda > 0$ and $\theta \in \mathbb{R}^d$. Consider the following optimization problem given by ridge regression:

$$\min_{\theta} \frac{1}{2}\|X\theta - y\|^2 + \frac{1}{2}\lambda\|\theta\|^2$$

For the following statements, answer whether they are TRUE or FALSE and provide a justification.

(a) Ridge regression can be viewed as logistic regression under a Bayesian perspective with a uniform prior on the parameters θ.

Response:

FALSE. Ridge regression can be viewed as linear regression under a Bayesian perspective with a Gaussian prior on θ.

Note $\log P(\theta) = -\lambda\|\theta\|^2 +$ constant in θ, for some λ

(b) Ridge regression has a unique solution if $\lambda > 0$, even if X has a null space.

Response:

TRUE

Solution is given by

$(X^tX + \lambda I)^{-1}X^ty$

As X^tX is positive semidefinite, its eigenvalues are nonnegative, so eigenvalues of $X^tX + \lambda I$ are all at least $\lambda > 0$. Thus $(X^tX + \lambda I)$ is invertible, and there is a unique point where

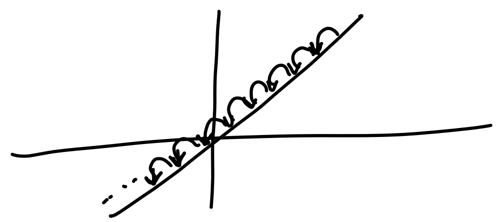
$\nabla_\theta \frac{1}{2}\|X\theta - y\|^2 + \frac{\lambda}{2}\|\theta\|^2 = 0.$

Question 5. *Gradient Descent*

Consider gradient descent with step size α on a function $f : \mathbb{R}^d \to \mathbb{R}$. Let $x^{(n)}$ be the nth iterate of gradient descent.

(a) TRUE or FALSE?
For any function f, if α is a small enough positive number, then $x^{(n)}$ will converge as $n \to \infty$. Provide a justification for your answer.

Response: False. If $f(x) = x$, then gradient descent for any $\alpha > 0$ will diverge to $-\infty$



(b) TRUE or FALSE? For a general function f, it is always the case that $f(x^{(n+1)}) < f(x^{(n)})$. If it is TRUE, provide a justification. If it is FALSE, present an example where this inequality does not hold and provide a justification.

Response:

False. If $f(x) = 0$, $x^{(n+1)} = x^{(n)}$, and so $f(x^{(n+1)}) = f(x^{(n)})$.

Question 6. *k Nearest Neighbors (KNN)*

(a) TRUE or FALSE? Using too small of a value of k for k-nearest neighbors would likely lead to overfitting. Provide a justification. **Response:**

TRUE. If $k=1$, for example, in a neighborhood of a point that contains noise, that noise will be reflected in the output of the predictor

(b) Describe a situation (in the context of regression) where using least squares linear regression would likely result in a better model than using KNN.

Response:

In a case where the true response is linear in the features.

KNN wont find a model that is linear in features (it is a piecewise constant function)

Question 7. *Linear Regression and Cross Validation*

Consider using linear regression with the following training data.

x	y
-1	-1
0	0
1	2

(a) Suppose you model the response $y = \theta_0 + \theta_1 x$. Using least squares linear regression, find the parameters θ_0, θ_1.

Response:

$$X = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \quad X^t X = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix} \quad (X^t X)^{-1} = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

$$\hat{\theta} = (X^t X)^{-1} X^t y = \begin{pmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix}$$

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

$$\boxed{\begin{aligned} \theta_0 &= \tfrac{1}{3} \\ \theta_1 &= \tfrac{3}{2} \end{aligned}}$$

(b) Using leave-one-out cross validation, estimate the test error of the predictor from part (a). Use the square loss to measure error.

Response:

Fold #	Train Set	Learned Model	holdout point	Square loss at holdout point
1	(0, 0) (1, 2)	$y = 0 + 2x$	(-1, -1)	$(-2+1)^2 = 1$
2	(-1, -1) (1, 2)	$y = \frac{1}{2} + \frac{3}{2}x$	(0, 0)	$(\frac{1}{2} - 0)^2 = \frac{1}{4}$
3	(-1, -1) (0, 0)	$y = 0 + x$	(1, 2)	$(1 - 2)^2 = 1$

So average square loss over the 3 folds is $\frac{1 + \frac{1}{4} + 1}{3} = \boxed{\frac{3}{4}}$