CS 189 Introduction to Machine Learning
Spring 2018 EXAMREVIEW1

1 Multiple Choice and Short Answer Questions

(a) What is the primary purpose of PCA?

◯ Dimension reduction ◯ Outlier removal

◯ Linear regression ◯ Optimization

Solution: (a) Dimension reduction is the primary purpose of PCA. While it can also be used for outlier removal, but that is a secondary purpose.

(b) Which of the following is **not** always true about a Multivariate Gaussian distribution?

◯ The isocontours are ellipses. ◯ Covariance matrix has positive entries.

◯ The PDF through the mean is Gaussian. ◯ The PDF parallel to an axis is Gaussian.

Solution: (c) The covariance matrix is positive semidefinite, but doesn't necessarily have positive entries. (Think of the matrix $\Sigma = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$.)

(c) Which of the following is true about training and testing error?

◯ Training measures bias; Testing measures variance. ◯ Training measures variance; Testing measures bias.

◯ Training measures bias; Testing measures both bias and variance. ◯ Training measures variance; Testing measures both bias and variance.

Solution: (b) Training measures bias; Testing measures both bias and variance. Recall the plots that we had in the bias-variance discussion, and in class. The training error will go to zero if and only if the data can be fit by a model in our class. It therefore measures bias. The test error is a sum of bias, variance, and irreducible error.

(d) When preforming polynomial regression, how does training error, validation error, and testing error change with the degree of the polynomial?

◯ Training Error decreases; Validation and Testing Error initial decrease, then increase.

◯ Training Error decreases; Validation and Testing Error decrease.

◯ Training Error initially decreases, then increases; Validation and Testing Error initial decrease, then increase;

◯ Training Error initially decreases, then increases; Validation and Testing Error decrease.

Solution: (a) Training Error decreases; Validation and Testing Error initial decrease, then increase.

(e) The left singular vectors of a rectangular matrix A are also:

◯ Eigenvectors of AA^T

◯ Eigenvectors of A^2

◯ Eigenvectors of A^TA

◯ Eigenvalues of AA^T

Solution: (a) Eigenvectors of AA^T. Straight from HW5, problem 3.

(f) The left singular vectors of a square matrix A are also:

◯ Eigenvectors of AA^T

◯ Eigenvectors of A^2

◯ Eigenvectors of A^TA

◯ Eigenvalues of AA^T

Solution: (a) Eigenvectors of AA^T. Straight from HW5, problem 3.

(g) Which of the following functions is **not** convex?

◯ $f(x) = e^{-x}$

◯ $f(x) = \sin x$

◯ $f(\mathbf{x}) = ||\mathbf{x}||_2^2$

◯ $f(x) = \max\{x, 0\}$

Solution: (c) $\sin(x)$ has second derivative $-\sin(x)$, which is not always positive.

2 Multiple choice, multiple may be correct

For these questions, multiple options could be correct. You will only get credit if you provide all the correct choices, and no partial credit.

(a) Assume you have two zero mean random variables $X \in \mathbb{R}^{d_1}$ and $Y \in \mathbb{R}^{d_2}$. Let their covariance matrices by given by $\Sigma_{XX}, \Sigma_{YY}, \Sigma_{XY}$. Also define the random variables $X' = A^\top X$, and $Y' = B^\top Y$, where $A \in \mathbb{R}^{d_1 \times d_1}$ and $B \in \mathbb{R}^{d_2 \times d_2}$ are both matrices. Let $W_1 \in \mathbb{R}^{d_1 \times d_1}$ and $W_2 \in \mathbb{R}^{d_2 \times d_2}$ represent two unitary matrices. Also, for any matrix M, we let $U(M)$ denote the matrix of its left singular vectors, and $V(M)$ denote the matrix of right singular vectors.

Which of the following is/are correct?

(a) Choosing $A = \Sigma_{XX}^{-1/2}$ whitens X'.

(b) Choosing $B = \Sigma_{XX}^{-1/2}$ whitens Y'.

(c) Choosing $B = W_2\Sigma_{YY}^{-1/2}$ whitens Y'.

(d) Choosing $A = U(\Sigma_{XX})$ decorrelates the entries of X'.

(e) Choosing $A = U(\Sigma_{XY})$ and $B = V(\Sigma_{XY})$ leads to diagonal $\Sigma_{X'Y'}$.

(f) If we want diagonal $\Sigma_{X'Y'}$ and whitened X', Y', using unitary matrices A, B is insufficient.

Solution: A white random variable has an identity covariance matrix. The covariances are given by $\mathbb{E}[X'X'^\top] = A^\top \Sigma_{XX} A$, $\mathbb{E}[Y'Y'^\top] = B^\top \Sigma_{YY} B$, $\mathbb{E}[X'Y'^\top] = A^\top \Sigma_{XY} B$.

Option (a) and (c) are thus correct, since we have $A^\top \Sigma_{XX} A = B^\top \Sigma_{YY} B = I$ in these cases. We can write the SVD $\Sigma_{XX} = U(\Sigma_{XX})\Lambda U(\Sigma_{XX})^\top$ to see that choosing $A = U(\Sigma_{XX})$ leads to the covariance $\mathbb{E}[X'X'^\top] = \Lambda$, thus correlating its entries. Option (d) is therefore correct.

Using a similar SVD, we see that option (e) is also correct. Option (f) is correct, since we saw that we need a matrix of the form $W_1\Sigma_{XX}^{-1/2}$ to whiten X', which is not necessarily unitary.

The correct options are therefore (a, c, d, e, f).

(b) Increasing λ in ridge regression has the following effect(s):

(a) The bias blows up without bound irrespective of the true model.

(b) The variance increases.

(c) The regularization strength increases.

(d) The model complexity decreases.

(e) The bias when $\lambda = 0$ is 0 if data is generated by a linear model.

(f) The bias when $\lambda = 0$ is always 0 irrespective of how the data is generated.

Solution: Increasing λ clearly increases the strength of regularization. At $\lambda = 0$, the estimated model has expectation w^* if the true model is linear, and at $\lambda \to \infty$, we have $\hat{w} = 0$, with zero variance. If the true model is linear and has $w^* = 0$, then the bias is always 0. The bias therefore increases, but not without bound for every true model. The variance decreases.

Note that we know that increasing λ has the interpretation of decreasing the norm of the allowed solution $\hat{w}$, called shrinkage (recall HW2). The model complexity (class of allowed models) therefore decreases with increasing λ.

If the true model is not linear, then at $\lambda = 0$, we incur the bias of all linear models.

Thus, (c, d, e) are true.

3 Estimation in linear regression

In linear regression, we estimate a vector $y \in \mathbb{R}^n$ by using the columns of a feature matrix $A \in \mathbb{R}^{n\times d}$. Assume that the number of training samples $n \geq d$ and that A has full column rank. You saw in homework how well we could predict y; let us now see how well we can estimate the regression coefficients.

Assume that the true underlying model for our noisy training observations is given by $Y = Aw^* + Z$, with $Z \in \mathbb{R}^n$ having iid $Z_j \sim \mathcal{N}(0,1)$ representing the random noise in the observation Y. Here, the $w^* \in \mathbb{R}^d$ is something arbitrary and not random. After obtaining $\hat{w} = \arg\min_w \|Y - Aw\|_2^2$, we would like to bound the error $\|\hat{w} - w^*\|_2^2$, which is our error in estimating the underlying parameters w^*. Note that this is a random variable.

Having a good estimate of the parameters is the ultimate goal, since we then know exactly how the underlying model is generated.

(a) Using the standard closed form solution to the ordinary least squares problem, **show that**

$$\|\hat{w} - w^*\|_2^2 = \|(A^\top A)^{-1} A^\top Z\|_2^2.$$

Solution: Note that the solution to the least squares problem is given by

$$\begin{aligned}\hat{w} &= (A^\top A)^{-1} A^\top y \\ &= (A^\top A)^{-1} A^\top A w^* + (A^\top A)^{-1} A^\top Z \\ &= w^* + (A^\top A)^{-1} A^\top Z.\end{aligned}$$

Hence, we have

$$\|\hat{w} - w^*\|_2^2 = \|(A^\top A)^{-1} A^\top Z\|_2^2,$$

which completes the proof.

(b) Use the (full) SVD of the matrix $A = U\Sigma V^\top$ to conclude that

$$\|\hat{w} - w^*\|_2^2 = \|V\Sigma' U^\top Z\|_2^2,$$

where we have denoted

$$\Sigma' = \begin{bmatrix} \Sigma_{\mathsf{inv}} & \mathbf{0}_{d\times(n-d)} \end{bmatrix}.$$

Here, we have used $\Sigma_{\mathsf{inv}} \in \mathbb{R}^{d\times d}$ to denote a diagonal matrix consisting of the reciprocals of the singular values of A, and $\mathbf{0}_{d\times(n-d)}$ denotes the $d \times (n-d)$ matrix of zeroes.

Solution: Given the SVD of A, notice that $(A^\top A)^{-1} = V\Sigma_{\mathsf{inv}}^2 V^\top$, and that $A^\top = V\Sigma^\top U^\top$. Using part (a), we have

$$\begin{aligned}\|\hat{w} - w^*\|_2^2 &= \|(A^\top A)^{-1} A^\top Z\|_2^2 \\ &= \|V\Sigma_{\mathsf{inv}}^2 V^\top V \Sigma^\top U^\top Z\|_2^2.\end{aligned}$$

For a matrix X with orthonormal columns, we have $X^\top X = I$. Additionally, we have $\Sigma_{\mathsf{inv}}^2 \Sigma^\top = \Sigma'$ (check this by simply multiplying these matrices), and so

$$\|\hat{w} - w^*\|_2^2 = \|V\Sigma' U^\top Z\|_2^2,$$

as desired.

(c) What is the distribution of $U^\top Z$? Use unitary invariance of the ℓ_2-norm and the distribution you calculated to conclude that

$$\|\hat{w} - w^*\|_2^2 = \|\Sigma' Z'\|_2^2,$$

where $Z' \in \mathbb{R}^n$ is also i.i.d. standard Gaussian.

Solution:

We know that a matrix times a standard Gaussian vector is jointly Gaussian, so it remains to calculate the mean and covariance. The mean of $Z' = U \top Z$ is clearly 0, and the covariance is given by

$$\mathbb{E}[Z' Z'^\top] = \mathbb{E}[U^\top Z Z^\top U] = U^\top \mathbb{E}[Z Z^\top] U = U^\top I U = I.$$

Hence, Z' is also i.i.d. standard Gaussian. Additionally, we know that $\|Vu\|_2 = \|u\|_2$ for any matrix V having orthnormal columns, and so combining these facts with the previous part, we have

$$\begin{aligned}
\|\hat{w} - w^*\|_2^2 &= \|V\Sigma' U^\top Z\|_2^2 \\
&= \|V\Sigma' Z'\|_2^2 \\
&= \|\Sigma' Z'\|_2^2.
\end{aligned}$$

(d) Now conclude that

$$\mathbb{E}\left[\|\hat{w} - w^*\|_2^2\right] = \mathsf{trace}[(A^\top A)^{-1}].$$

Which of the following matrices A is better for estimation of the parameters?

i) $A_1 = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$

ii) $A_2 = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}$

Solution: We know from the previous part that

$$\begin{aligned}
\|\hat{w} - w^*\|_2^2 &= \|\Sigma' Z'\|_2^2 \\
&= \sum_{i=1}^{d} \left(\frac{Z_i'}{\sigma_i}\right)^2,
\end{aligned}$$

where we have used σ_i to denote the ith singular value of the matrix A. Now taking the expectation, we know that since $\mathbb{E}[(Z_i')^2] = 1$, we have

$$\begin{aligned}
\mathbb{E}[\|\hat{w} - w^*\|_2^2] &= \sum_{i=1}^{d} \frac{1}{\sigma_i^2} \\
&= \mathsf{trace}[(A^\top A)^{-1}],
\end{aligned}$$

where we have used the fact that the ith eigenvalue of $(A^\top A)^{-1}$ is given by $1/\sigma_i^2$.

Since $tr((A_1^T A_1)^-1) = \frac{1}{25} + \frac{1}{25} = \frac{2}{25}$ and $tr((A_2^T A_2)^-1) = \frac{1}{100} + \frac{1}{4} = \frac{26}{100}$, it is clear that A_1 would be a better estimation of the parameters.

4 (NOT IN SCOPE FOR SP18 MIDTERM EXAM) Convergence Rate of Gradient Descent for Quadratic Functions

Show that the following problems have a geometric convergence rate when applying gradient descent with a fixed step size. Recall from Homework 6, Problem 3 that for a constant step size $\gamma = \frac{1}{\lambda_{\min}(A^TA)+\lambda_{\max}(A^TA)}$ and A^TA positive definite,

$$\min_x \frac{1}{2}\|Ax-b\|_2^2$$

has geometric convergence. That is, for $Q = \frac{\lambda_{\max}(A^TA)}{\lambda_{\min}(A^TA)}$, we have

$$f(x_k) - f(x^*) = \frac{\lambda_{\max}(A^TA)}{2}\left(\frac{Q-1}{Q+1}\right)^{2k}\|x_0 - x^*\|_2^2$$

You may use the above result (without rederivation) if required for the following parts.

(a) Consider a matrix $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n, k \in \mathbb{R}$ such that $A^TA \succeq mI$ for $m > 0$:

$$\min_x \frac{1}{2}\|Ax-b\|_2^2 + k$$

Show geometric convergence for this problem.

Solution: Adding a constant changes neither the gradient of the function at any particular point x nor the optimal solution x^* (although it may change the optimal objective value). Therefore, one can solve instead the problem without the added constant

$$\min_x \frac{1}{2}\|Ax-b\|_2^2$$

Then, by HW3 Problem 6, geometric convergence is attained.

(b) Consider $A \in \mathbb{R}^{n\times n}$ (a square matrix) and $b, c \in \mathbb{R}^n$, such that $A \succeq mI$ for $m > 0$:

$$\min_x \frac{1}{2}\|Ax-b\|_2^2 + c^Tx$$

Show geometric convergence for this problem.

Solution: We can formulate the problem as

$$\min_x \frac{1}{2}\|A'x-b'\|_2^2$$

where

$$A' = A \text{ and } b' = b - \frac{1}{2}c^TA^{-1}$$

Expanding out, we see

$$\begin{aligned}
&\min_x \frac{1}{2} x A'^T A' x - 2b'^T A' x + \frac{1}{2} b'^T b' \\
&= \frac{1}{2} x A^T A x - 2b^T A' x + c^T A^{-1} A x + \frac{1}{2} b'^T b' \\
&= \frac{1}{2} \|Ax - b\|_2^2 + c^T x
\end{aligned}$$

We observe that $A'^T A' = A^T A \succeq m > 0$. That is, we reduce the given problem to the quadratic problem above. Then, by HW3 Problem 6, geometric convergence is attained.

(c) Consider $A, C \in \mathbb{R}^{n \times d}, b, d \in \mathbb{R}^n$ such that $A^T A \succeq m_1 I, C^T C \succeq m_2 I$ for $m_1, m_2 > 0$:

$$\min_x \frac{1}{2} \|Ax - b\|_2^2 + \frac{1}{2} \|Cx - d\|_2^2$$

Show geometric convergence for this problem.

Solution: We can formulate the problem as

$$\min_x \frac{1}{2} \|A'x - b'\|_2^2$$

where

$$A' = \begin{bmatrix} A \\ C \end{bmatrix} \text{ and } b' = \begin{bmatrix} b \\ d \end{bmatrix}$$

We observe that $A'^T A' = A^T A + C^T C \succeq m_1 + m_2 > 0$. That is, we reduce the given problem to the quadratic problem above. Then, by HW3 Problem 6, geometric convergence is attained.

5 Parameter Estimation

Assume that $X_1, X_2, \ldots, X_n$ are i.i.d. samples from an exponential distribution $p(X = x) = \lambda e^{-\lambda x}$.

(a) Compute the maximum likelihood estimation of λ given $X_1, \ldots, X_n$.

Solution:

$$l(\lambda) = \Sigma_{i=1}^n (\log \lambda - \lambda X_i) = n \log \lambda - \lambda \Sigma_{i=1}^n X_i$$

Set the derivative to zero we have:

$$\begin{aligned}
\frac{n}{\lambda} - \Sigma_{i=1}^n &= 0 \\
\lambda &= \frac{n}{\Sigma_{i=1}^n X_i}
\end{aligned}$$

(b) Now assume the the prior on λ has a gamma distribution with parameters α, β:

$$P(X = x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

where $\Gamma(\alpha)$ is the gamma function, which you should think of as a constant normalization that ensures that the PDF integrates to 1.

Show that the posterior random variable $\lambda | (X_1, X_2, \ldots, X_n)$ also has a gamma distribution.

Solution: Define $X = X_1, ..., X_n$.

$$\begin{aligned} p(\lambda|X) &\propto P(X|\lambda)P(\lambda) \\ &\propto \lambda^n e^{-\lambda \Sigma_{i=1}^n X_i} \lambda^{\alpha-1} e^{-\beta\lambda} \\ &\propto \lambda^{n+\alpha-1} e^{-(\Sigma_{i=1}^n X_i + \beta)\lambda} \end{aligned}$$

The parameters of the gamma distribution are $\alpha + n, \Sigma_{i=1}^n X_i + \beta$.

(c) Compute the maximum a posteriori estimation of λ. Compare the result with the maximum likelihood estimation of λ when the sample size is large.

Solution:

We set the derivative of log posterior to zero:

$$\log(P(\lambda|X) \propto (n + \alpha - 1)\log\lambda - (\Sigma_{i=1}^n X_i + \beta)\lambda$$

$$\frac{d(\log(P(\lambda|X))}{d\lambda} = \frac{(n+\alpha-1)}{\lambda} - (\Sigma_{i=1}^n X_i + \beta) = 0$$

$$\lambda = \frac{(n+\alpha-1)}{\Sigma_{i=1}^n X_i + \beta}$$

If n is large the MAP and MLE estimation will be the same. Having a good prior on parameters help when the sample size is small.

(d) Now assume that $X_1, X_2, ..., X_n$ are i.i.d. samples from a gamma distribution $P(X = x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$. In the next parts we will compute the gradient descent step for the maximum likelihood estimation of α and β. Assume that

$$\frac{d\Gamma(x)}{dx} = g(x)$$

Compute the partial derivative of log likelihood with respect to β and find β as a function of α.

Solution:

$$l(\alpha,\beta) = n\alpha\log\beta - n\log\Gamma(\alpha) + (\alpha-1)\Sigma_{i=1}^{n}\log X_i - \beta\Sigma_{i=1}^{n} X_i$$

Take the partial derivative of log likelihood with respect to β

$$\frac{dl}{d\beta} = \frac{n\alpha}{\beta} - \Sigma_{i=1}^{n} X_i$$

We can directly compute β as a function of α if we put this gradient to zero.

$$\beta = \frac{n\alpha}{\Sigma_{i=1}^{n} X_i} = \frac{\alpha}{\mu_x}$$

(e) Using your result for the previous part, compute the gradient descent step for α.

Solution:

We substitute $\beta = \frac{\alpha}{\mu_x}$ in log likelihood and then we compute the partial derivative of log likelihood with respect to α

Gradient descent step for α will be:

$$\frac{dl}{d\alpha} = n\log\alpha - n\log\mu_x + n - n\frac{g(\alpha)}{\Gamma(\alpha)} + \Sigma_{i=1}^{n}\log X_i$$

