CS 189 Fall 2016 Introduction to Machine Learning Final

- Do not open the exam before you are instructed to do so.
- The exam is closed book, closed notes except your one-page cheat sheet.
- Usage of electronic devices is forbidden. If we see you using an electronic device (phone, laptop, etc.) you will get a zero.
- You have 3 hours.
- Write your initials at the top right of each page (e.g., write "BR" if you are Ben Recht).
- Mark your answers on the exam itself in the space provided. Do not attach any extra sheets.

First name	
Last name	
SID	
First and last name of student to your left	
First and last name of student to your right	

Q1. [10 pts] Inverse Gaussian Distribution

The inverse Gaussian distribution, denoted $\text{IG}(\mu, \lambda)$, is a continuous distribution supported on $(0, \infty)$ and parameterized by two scalars $\mu > 0$ and $\lambda > 0$. For any $x \in (0, \infty)$, the probability density function $f(x; \mu, \lambda)$ is given by

$$f(x; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\} .$$

(a) [3 pts] Suppose we draw n independent samples $x_1, ..., x_n$ from $\text{IG}(\mu, \lambda)$. Write down the log-likelihood $\mathcal{L}(x_1, ..., x_n; \mu, \lambda)$.

We have that for $x \in (0, \infty)$,

$$\log f(x; \mu, \lambda) = \frac{1}{2}\log\lambda - \frac{1}{2}\log 2\pi x^3 - \frac{\lambda(x-\mu)^2}{2\mu^2 x} .$$

Hence,

$$\mathcal{L}(x_1, ..., x_n; \mu, \lambda) = \sum_{i=1}^{n} \log f(x_i; \mu, \lambda) = \frac{n}{2}\log\lambda - \sum_{i=1}^{n}\frac{1}{2}\log 2\pi x_i^3 - \sum_{i=1}^{n}\frac{\lambda(x_i-\mu)^2}{2\mu^2 x_i} .$$

(b) [2 pts] Is the function $\phi : (0, \infty) \longrightarrow \mathbb{R}$ defined as $\phi(\lambda) = \mathcal{L}(x_1, ..., x_n; \mu, \lambda)$ convex, concave, or both? No need to justify your answer.

Concave. $\phi(\lambda) = a\log\lambda + b\lambda + c$ for scalars a, b, c with $a > 0$, and $\log\lambda$ is concave on $(0, \infty)$.

(c) [5 pts] Now assume that μ is known. Derive the maximum likelihood estimator $\widehat{\lambda}$ given n independent samples $x_1, ..., x_n$ from $\text{IG}(\mu, \lambda)$.

Solving for the stationary point of $\frac{d}{d\lambda}\mathcal{L}(x_1, ..., x_n; \mu, \lambda) = 0$,

$$0 = \frac{d}{d\lambda}\mathcal{L}(x_1, ..., x_n; \mu, \lambda) = \frac{n}{2\lambda} - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\mu^2 x_i} \Longrightarrow \widehat{\lambda}^{-1} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\mu^2 x_i} .$$

Since the function $\phi(\lambda)$ is concave on $(0, \lambda)$ and the proposed maximizer is positive with probablity one, then the solution to $\frac{d}{d\lambda}\mathcal{L}(x_1, ..., x_n; \mu, \lambda) = 0$ is both a necessary and sufficient condition for a global maximum.

Q2. [25 pts] Multivariate Gaussians

Let x be a d-dimensional random vector distributed as a Gaussian with mean $\mu \in \mathbb{R}^d$ and positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$: that is, $x \sim N(\mu, \Sigma)$.

(a) [5 pts] Let $v \in \mathbb{R}^d$. What are the mean and variance of $v^\mathsf{T} x$? $v^\mathsf{T}\mu$ and $v^\mathsf{T}\Sigma v$.

(b) [5 pts] What is the distribution of $v^\mathsf{T} x$? $N(v^\mathsf{T}\mu, v^\mathsf{T}\Sigma v)$.

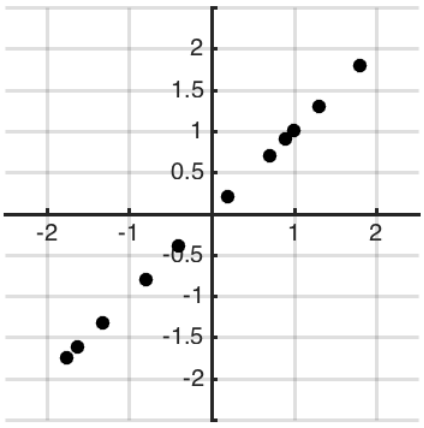
(c) [5 pts] Suppose we get to choose v from the unit sphere (i.e. $\|v\|_2 = 1$). Write down a (unit-norm) choice of v that maximizes the variance of $v^\mathsf{T} x$. Leading eigenvector of Σ.

(d) [5 pts] Let $z \sim N(0, I_d)$. Find a matrix A and a vector b such that $Az + b$ has the same distribution as x. $A = \Sigma^{1/2}, b = \mu$.

(e) [5 pts] Now find a matrix A and a vector b such that $Ax + b$ has the same distribution as z. $A = \Sigma^{-1/2}, b = -\Sigma^{-1/2}\mu$.

Q3. [20 pts] Data Geometry

Consider the following two-dimensional data set:



Note that the mean of this data set is 0.

(a) [5 pts] Suppose someone assigns a label $y_i = +1$ or $y_i = -1$ to every data point and we solve a Support Vector Machine (SVM) classification problem:

$$\text{minimize}_{w \in \mathbb{R}^2, b \in \mathbb{R}} \quad \frac{1}{n} \sum_{i=1}^{n} \max(1 - y_i(w^\top x_i + b), 0) + \lambda \|w\|_2^2 \,.$$

Here, b is an unregularized bias term, and $\lambda > 0$. Show that both components of w must be equal to each other.

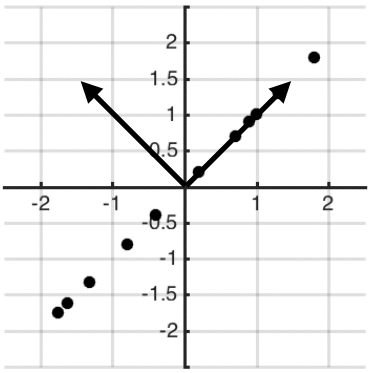
The optimal $w_\star$ must lie in the span of the data. Therefore, since all of the data points have both components equal to each other, $w_\star$ will as well.

(b) [5 pts] Suppose $\lambda = 0$. Given the information provided, what can be said about the possible values for w?

The optimal $w_\star$ could be any direction in $\mathbb{R}^2$ based on the information provided. Without regularization, w can roam free.

(c) [5 pts] Draw the principal components of this data set on the plot.

Solution:

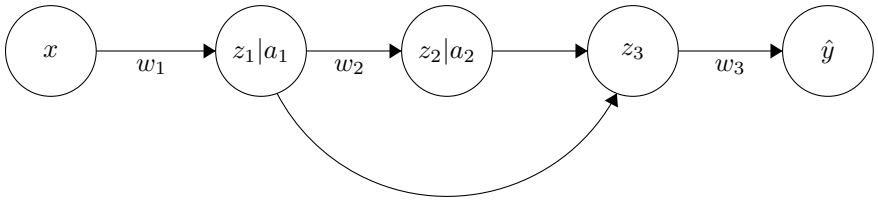


(d) [5 pts] What is the ratio of the smallest eigenvalue to the largest eigenvalue of the covariance matrix of this data?

0.

Q4. [20 pts] Residual Neural Network

Consider the following neural network, which operates on scalars.



In this network, w_1, w_2, and w_3 are scalars. The network takes the scalar x as input, and computes $z_1 = w_1 x$. The ReLU nonlinearity is then applied: $a_1 = \text{ReLU}(z_1)$. Next, the network computes $z_2 = w_2 a_1$ and applies the ReLU nonlinearity: $a_2 = \text{ReLU}(z_2)$. Next, $z_3 = a_1 + a_2$. Finally, $\hat{y} = w_3 z_3$.

We will use the mean squared error loss function $L = \frac{1}{2}(y - \hat{y})^2$ to train this network. You may use $R(x)$ and $R'(x)$ to denote ReLU and the derivative of ReLU respectively.

(a) [5 pts] Find $\frac{dL}{dw_3}$.

$$\frac{dL}{dw_3} = -(y - \hat{y}) z_3$$

(b) [5 pts] Find $\frac{dL}{dw_2}$.

$$\frac{dL}{dw_2} = -(y - \hat{y}) w_3 \frac{dR(w_2 a_1)}{dw_2} = -(y - \hat{y}) w_3 R'(z_2) a_1$$

(c) [5 pts] Find $\frac{dL}{dw_1}$.

$$\begin{aligned}
\frac{dL}{dw_1} &= -(y - \hat{y}) w_3 \frac{d(R(w_1 x) + R(w_2 R(w_1 x)))}{dw_1} \\
&= -(y - \hat{y}) w_3 (R'(w_1 x) x + R'(w_2 R(w_1 x)) w_2 R'(w_1 x) x) \\
&= -(y - \hat{y}) w_3 R'(z_1) x (1 + R'(z_2) w_2)
\end{aligned}$$

(d) [5 pts] How would the network change if we added a ReLU nonlinearity to unit z_3 such that $a_3 = \text{ReLU}(z_3), \hat{y} = w_3 a_3$. *Briefly* explain your reasoning.

We already have that $a_1 \geq 0, a_2 \geq 0$, so it would be redundant to apply ReLU to z_3.

Q5. [35 pts] Feature Engineering

In this class, we have tried to emphasize the importance of good features. In this question, we'll take this a step further and see that good features need to be predictive on the training set, but also generalize well to unseen data.

Let n denote the number of UC Berkeley students, and let us assign an arbitrary ordering to all students. Let $y_1, y_2, ..., y_n \in \mathbb{R}$ denote the ages of all the students. Our goal is to predict the age y of a new transfer student on campus based on some list of features. A natural feature choice could be, for example, year of study. But suppose we only have access to student ID (SID) numbers. You are a bit unsure how to use this information, so you ask some of your friends.

Note: For this question, assume that $n < 10^9$.

One-hot encoding. You first ask your friend from Stanford how to encode a student ID into a feature vector. "Use one-hot encoding", your friend responds. You are a bit skeptical, but you proceed onwards. Since there are 9 digits in an SID, you one-hot encode each SID into an element of $\{0,1\}^d$ where $d = 10^9$. Let $x_1, ..., x_n \in \mathbb{R}^d$ denote the encoded features of the training set stacked row-wise into a matrix $X \in \mathbb{R}^{n \times d}$, and $Y = (y_1, ..., y_n) \in \mathbb{R}^n$ denote the age vector of students in the training set (note that $n < d$). You proceed to use regularized least squares and solve

$$\text{minimize}_{w \in \mathbb{R}^d} \ \frac{1}{2}\|Xw - Y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 \,. \tag{1}$$

(a) [3 pts] How does the regularization term ($\frac{\lambda}{2}\|w\|_2^2$) ensure a unique solution to (1)?

The matrix X is of dimension $n \times 10^9$ with $n < 10^9$, so the matrix X has a non-trivial null-space.

(b) [5 pts] Compute the solution $w_\star \in \mathbb{R}^{10^9}$ to (1), assuming that $x_i = e_i \in \mathbb{R}^d$ where e_i denotes the i-th standard basis vector in $\mathbb{R}^d$. That is, someone sorted your data such that the i-th student has SID i. Your answer should only be expressed in terms of Y and λ.

Hint: You may perform this calculation either directly, or using the kernel trick.

(Direct.) We know that $w_\star = (X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}Y$. Hence,

$$X = \begin{bmatrix} I_n & 0_{n,d-n} \end{bmatrix} \,, \ X^\mathsf{T}X = \sum_{i=1}^{n} e_i e_i^\mathsf{T} = \begin{bmatrix} I_n & 0_{n,d-n} \\ 0_{d-n,n} & 0_{d-n,d-n} \end{bmatrix} .$$

Therefore,

$$\begin{aligned} w_\star = (X^\mathsf{T}X + \lambda I_n)^{-1}X^\mathsf{T}Y &= \begin{bmatrix} \frac{1}{1+\lambda} I_n & 0_{n,d-n} \\ 0_{d-n,n} & \frac{1}{\lambda} I_{d-n,d-n} \end{bmatrix} \begin{bmatrix} I_n \\ 0_{d-n,n} \end{bmatrix} Y \\ &= \begin{bmatrix} \frac{1}{1+\lambda} I_m \\ 0_{d-n,n} \end{bmatrix} Y = \begin{bmatrix} \frac{1}{1+\lambda} Y \\ 0_{d-n} \end{bmatrix} . \end{aligned}$$

(Kernel trick.) It is not hard to see that the gram matrix $XX^\mathsf{T} = I_n$. Recall that $w_\star = X^\mathsf{T}\alpha_\star$, and $\alpha_\star = (XX^\mathsf{T} + \lambda I_n)^{-1}Y$. Clearly then $\alpha_\star = \frac{1}{1+\lambda}Y$ and $w_\star = X^\mathsf{T}\alpha_\star = \begin{bmatrix} \frac{1}{1+\lambda}Y \\ 0_{d-n} \end{bmatrix}$.

(c) [3 pts] Compute $\text{predict}(x_i) = \langle w_\star, x_i \rangle$ for x_i in the training set.

$$\text{predict}(x_i) = \frac{1}{1+\lambda} y_i \,.$$

(d) [3 pts] Suppose x is the feature vector for the transfer student who is *not* in the training set. Compute $\text{predict}(x) = \langle w_\star, x \rangle$ for x not in the training set.

Since x is not in the training set, its one-hot encoding is orthogonal to every feature vector in the training set. Thus

$$\text{predict}(x) = 0\,.$$

Logical features. Unsatisfied with the performance of one-hot encoding, you ask another friend from MIT what features to use. Your friend tells you to use the feature that is equal to 1 if the student goes to Berkeley, and 0 otherwise. You are even more skeptical, but you proceed onwards.

(e) [5 pts] Using the features $x_i = 1$ for all $1 \le i \le n$, write the solution $w_\star$ to the following *unregularized* problem:

$$\text{minimize}_{w \in \mathbb{R}} \ \frac{1}{2}\|Xw - Y\|_2^2\,.$$

Our features in this case are scalars. We know that $X = \mathbf{1}_n$ and hence $X^\top X = \mathbf{1}_n^\top \mathbf{1}_n = n$. The solution is simply $w_\star = \frac{1}{n}\sum_{i=1}^n y_i$.

(f) [3 pts] Now compute $\text{predict}(x_i) = \langle w_\star, x_i \rangle$ using the $w_\star$ from part (e) for x_i in the training set.

$$\text{predict}(x_i) = \frac{1}{n}\sum_{i=1}^n y_i\,.$$

(g) [3 pts] Since the transfer student is a Berkeley student, the value of their feature will be $x = 1$. Compute $\text{predict}(x) = \langle w_\star, x \rangle$ using the $w_\star$ when x is the transfer student who is *not* in the training set.

Again,

$$\text{predict}(x_i) = \frac{1}{n}\sum_{i=1}^n y_i\,.$$

Combining features. Still unsatisfied with previous suggestions, we consider combining both the one-hot encoded features and the constant feature.

(h) [10 pts] We now consider $x_i = (e_i, 1) \in \mathbb{R}^{d+1}$, where $e_i \in \mathbb{R}^d$ as before. For simplicity, we set $\lambda = 0$, and consider the following problem.

$$\text{minimize}_{w \in \mathbb{R}^{d+1}} \ \frac{1}{2}\|Xw - Y\|_2^2 \,.$$

Compute a minimizer $w_\star$ of this problem, and predict(x) for x in and not in the training set.

Hint: The Sherman-Morrison formula states that for invertible A and column vectors u, v such that $1 + v^\mathsf{T} A^{-1} u \neq 0$,

$$(A + uv^\mathsf{T})^{-1} = A^{-1} - \frac{A^{-1}uv^\mathsf{T}A^{-1}}{1 + v^\mathsf{T}A^{-1}u} \,.$$

Since we are allowed to compute any minimizer, it is easiest to compute the minimum norm one by employing the kernel trick. Observe that

$$X = \begin{bmatrix} I_n & 0_{n,d-n} & \mathbf{1}_n \end{bmatrix} \,, \ XX^\mathsf{T} = I_n + \mathbf{1}_n \mathbf{1}_n^\mathsf{T} \,.$$

By the Sherman-Morrison formula, we have that

$$(XX^\mathsf{T})^{-1} = (I_n + \mathbf{1}_n\mathbf{1}_n^\mathsf{T})^{-1} = I_n - \frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{1+n} \,.$$

Hence,

$$\alpha_\star = (XX^\mathsf{T})^{-1}Y = \left(I_n - \frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{1+n}\right) Y \,.$$

Furthermore, the minimum norm minimizer $w_\star$ is given by

$$w_\star = X^\mathsf{T}\alpha_\star = \begin{bmatrix} I_n \\ 0_{d-n,n} \\ \mathbf{1}_n^\mathsf{T} \end{bmatrix} \left(I_n - \frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{1+n}\right) Y = \begin{bmatrix} \left(I_n - \frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{1+n}\right) Y \\ 0_{d-n} \\ \mathbf{1}_n^\mathsf{T}Y - \frac{n}{n+1}\mathbf{1}_n^\mathsf{T}Y \end{bmatrix} = \begin{bmatrix} \left(I_n - \frac{\mathbf{1}_n\mathbf{1}_n^\mathsf{T}}{1+n}\right) Y \\ 0_{d-n} \\ \frac{1}{n+1}\sum_{i=1}^n y_i \end{bmatrix} \,.$$

We are now in a position to compute the predict(x) function. On the training set.

$$\text{predict}(x_i) = y_i \,.$$

For x not in the training set,

$$\text{predict}(x) = \frac{1}{n+1}\sum_{i=1}^n y_i \,.$$

Q6. [20 pts] Calibration

Suppose we attempt to solve a binary classification task with binary feature vectors. Specifically, we are given n data points $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \{0,1\}^d$ (i.e. each of the d feature values is either 0 or 1) and $y_i \in \{0,1\}$. Suppose we try to fit the standard linear logistic regression model:

$$P(y = 1|x, w, b) = g(w^\mathsf{T} x + b) = \frac{1}{1 + \exp(-w^\mathsf{T} x - b)} .$$

We say that a model is *calibrated* if

$$\frac{1}{n}\sum_{i=1}^n P(y_i = 1|x_i, w, b) = \frac{1}{n}\sum_{i=1}^n y_i .$$

That is, if the average probability of labeling the data in class 1 is equal to the average number of occurrences of class 1 on the training set. Calibrated models are interesting because their predicted probabilities are more useful as an indicator of confidence.

Consider training the logistic regression model where we solve the optimization problem

$$\text{maximize}_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad \sum_{i=1}^n \log P(y_i|x_i, w, b) .$$

Let $w_\star, b_\star$ denote the maximizing parameters.

(a) [5 pts] Show that the logistic model corresponding to the optimal $w_\star, b_\star$ is calibrated.

Note that the log-likelihood can be written as

$$\sum_{i=1}^n y_i(w^\mathsf{T} x_i + b) - \log(1 + \exp(w^\mathsf{T} x_i + b))$$

Taking a derivative with respect to b shows that $b_\star$ must satisfy

$$0 = \sum_{i=1}^n \left\{ y_i - \frac{\exp(w_\star^\mathsf{T} x_i + b_\star)}{1 + \exp(w_\star^\mathsf{T} x_i + b_\star)} \right\} = \sum_{i=1}^n y_i - \sum_{i=1}^n P(y_i = 1|x_i, w_\star, b_\star) .$$

Another (simpler) method:

Denote $\mathcal{L}$ as the log-likelihood, $\mu_i = P(y_i = 1|x_i, w, b)$ Reformulate $w^T x_i + b$ as $\beta^T x_i$, with $\beta = [w, 1]^T$, $x_i := [x_i, 1]^T \in \mathbb{R}^{d+1}$. From lecture/homework we know:

$$\nabla_\beta \mathcal{L} = \sum_i^n (y_i - \mu_i)x_i = 0$$

Noting that $x_{i,d} = 1$ (the bias component added to the end of x), we have:

$$\sum_i^n y_i - \mu_i = 0 \implies \sum_i^n y_i = \sum_i^n \mu_i$$

Many students made the mistake of claiming that under the optimal $w_\star, b_\star$, $P(y_i|x_i, w, b) = y_i$. This is *only* true if the data is linearly separable, and even then it is only true as $||w|| \to \infty$.

(b) [5 pts] Let S_j denote the set of indices i where $X_{ij} = 1$. We say a model is *conditionally* calibrated with respect to feature j if

$$\frac{1}{|S_j|}\sum_{i\in S_j} P(y_i = 1|x_i, w, b) = \frac{1}{|S_j|}\sum_{i\in S_j}^n y_i .$$

Show that the logistic model corresponding to the optimal $w_\star, b_\star$ is calibrated with respect to every feature.

Taking a derivative with respect to w_j shows that $w_{\star,j}$ must satisfy

$$0 = \sum_{i=1}^{n} \left\{ y_i x_{ij} - \frac{x_{ij} \exp(w_\star^\mathsf{T} x_i + b_\star)}{1 + \exp(w_\star^\mathsf{T} x_i + b_\star)} \right\} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} P(y_i = 1 | x_i, w_\star, b_\star)$$
$$= \sum_{i \in S_j} y_i - \sum_{i \in S_j} P(y_i = 1 | x_i, w_\star, b_\star) \,.$$

Here, the last line follows because $x_{ij} = 1$ only if $i \in S_j$, and otherwise $x_{ij} = 0$.

Another (simpler) method:

Using the same notation as the simpler method in the previous part, we see that:

$$\sum_{i}^{n} (y_i - \mu_i) x_i = 0 \implies \sum_{i}^{n} (y_i - \mu_i) x_{ij} = 0, \forall j \implies \sum_{i \in S_j} y_i - \mu_i = 0$$

(c) [5 pts] Suppose we add a regularizer for the w parameters to the optimization problem:

$$\text{maximize}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \sum_{i=1}^{n} \log P(y_i | x_i, w, b) + \lambda \|w\|_2^2 \,.$$

Is the logistic model corresponding to the optimal $w_\star, b_\star$ calibrated? Is it calibrated with respect to every feature?

The model remains calibrated, as the gradient for b does not change, and hence the same calculation as in part (a) applies. However, for w, we now have

$$\frac{\partial \text{cost}}{\partial w_j} = \sum_{i \in S_j} y_i - \sum_{i \in S_j} P(y_i = 1 | x_i, w_\star, b_\star) + w_j$$

which will not be calibrated with respect to feature j when $w_j \neq 0$.

(d) [5 pts] Suppose someone gives you a classifying function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ that inputs x and outputs a real number. We can produce a probability from this classifier by setting

$$P(y = 1 | x, f, \alpha, \beta) = \frac{1}{1 + \exp(\alpha f(x) + \beta)}$$

Show that one can always find values α and β such that the resulting probability estimates are calibrated.

If one fits the log likelihood over the data, the optimality conditions with respect to β will guarantee calibration.

Another way to think about this: you can think of f as essentially a feature transformation. Replace your entire training set with $f(x)$. Then train logistic regression on this training set. The optimality conditions we already proved guarantee calibration.