Mathematics for Machine Learning
Problem Set 5
COMS 4995-002 : Summer B 2024

Instructor: Samuel Deng
samdeng@cs.columbia.edu
Due: 08/13/2024 11:59 PM ET.

Problem 1

Properties of random vectors [25 points].

In this problem, you will get a little more familiar with *random vectors* and their properties. Random vectors generalize random variables to multiple dimensions, but they aren't much harder to reason about than random variables.

A random vector in $\mathbb{R}^d$ is simply a vector of random variables. For random variables $x_1, \ldots, x_d$, we will write the corresponding random vector as:

$$\mathbf{x} := \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix},$$

just as we did for regular vectors, but we will simply keep in mind that these objects are random variables. The random variables that make up the coordinates of a random vector are not necessarily independent. The *expectation* of a random vector is nicely defined just as the coordinate-wise expectation. That is,

$$\mathbb{E}[\mathbf{x}] = \begin{bmatrix} \mathbb{E}[x_1] \\ \vdots \\ \mathbb{E}[x_d] \end{bmatrix}.$$

First, we will make sure that the familiar properties of expectation generalize over to random vectors. The most basic and useful property of expectation is linearity. For any scalar $\alpha \in \mathbb{R}$ and random variables X and Y, the expectation obeys:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \quad \text{and} \quad \mathbb{E}[\alpha X] = \alpha \mathbb{E}[X].$$

Problem 1(a) [3 points] Show that the linearity of expectation holds for random vectors. That is, for a fixed scalar $\alpha \in \mathbb{R}$ and random vector $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbb{E}[\alpha \mathbf{x}] = \alpha \mathbb{E}[\mathbf{x}].$$

Also, show that for any two random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\mathbb{E}[\mathbf{x} + \mathbf{y}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{y}].$$

Another familiar property of expectation is that, for *independent* random variables X and Y, the expectation of their product can be factored:

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \quad \text{if } X, Y \text{ are independent.}$$

Problem 1(b) [3 points] Let $\mathbf{x} = (x_1, \ldots, x_d)$ and $\mathbf{y} = (y_1, \ldots, y_d)$ be random vectors in $\mathbb{R}^d$. Prove that if they are element-wise independent, i.e. x_i is independent of y_i for all $i = 1, \ldots, d$, then:

$$\mathbb{E}[\mathbf{x}^\top \mathbf{y}] = (\mathbb{E}[\mathbf{x}])^\top (\mathbb{E}[\mathbf{y}]).$$

We also generalize the notion of *variance* of random variables to random vectors. Recall that the variance of a random variable X is given by its "average squared distance" from its mean:

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2].$$

For a pair of random variables X, Y, a related notion is the *covariance*, which we learned quantifies the "average linear relationship" between two random variables:

$$\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Observe that the covariance of a random variable and itself is simply its variance: $\mathrm{Cov}(X, X) = \mathrm{Var}(X, X)$. We define the variance of a random vector as a matrix, the random vector's *covariance matrix*.[1] Wedenote this as $\mathrm{Var}(\mathbf{x})$ or $\mathrm{Cov}(\mathbf{x})$, defined as:

$$\mathrm{Cov}(\mathbf{x}) = \mathrm{Var}(\mathbf{x}) := \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top].$$

Notice that, because $\mathbf{x} - \mathbb{E}[\mathbf{x}]$ is a d-dimensional vector, this gives us a $d \times d$ symmetric matrix.

Problem 1(c) [4 points] Recall that variance could also be written with the following equality:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

For any random vector $\mathbf{x} \in \mathbb{R}^d$, prove the property that:

$$\mathrm{Var}(\mathbf{x}) = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top.$$

[1] Note that, in other texts, what we are calling the "variance" of a random vector might be defined as the *covariance* of a random vector, $\mathrm{Cov}(\mathbf{x})$, and variance is defined separately, as a notion of "length." We will not consider that notion in this course, so we simply write $\mathrm{Var}(\mathbf{x})$ and $\mathrm{Cov}(\mathbf{x})$, to mean the random vector's covariance matrix.

Problem 1(d) [4 points] We can think of the covariance matrix of a random vector as a matrix containing all pair-wise covariances of its constituent random variables entries.

For an arbitrary random vector $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, show that the corresponding covariance matrix is given by:

$$\mathrm{Var}(\mathbf{x}) = \begin{bmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) & \ldots & \mathrm{Cov}(x_1, x_d) \\ \mathrm{Cov}(x_2, x_1) & \mathrm{Var}(x_2) & \ldots & \mathrm{Cov}(x_2, x_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_d, x_1) & \mathrm{Cov}(x_d, x_2) & \ldots & \mathrm{Var}(x_d) \end{bmatrix}.$$

That is, show that entry i, j of the matrix $\mathrm{Var}(\mathbf{x})$ is $\mathrm{Cov}(x_i, x_j)$:

$$[\mathrm{Var}(\mathbf{x})]_{ij} = \mathrm{Cov}(x_i, x_j).$$

State why this implies that the covariance matrix is symmetric.

Typically, the covariance matrix of a random vector $\mathrm{Var}(\mathbf{x})$ is denoted as $\Sigma \in \mathbb{R}^d$, when the context is clear. In Problem 1(d), we showed that this matrix is symmetric, so all the usual tools of linear algebra for symmetric matrices are at our disposal. In particular, we may reach for the spectral theorem to analyze the eigendecomposition of such matrices

$$\Sigma = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top,$$

as you will see in our final concerning the multivariate Gaussian distribution. It turns out that Σ is not only symmetric, but it is also positive semidefinite.

Problem 1(e) [4 points] Let $\mathbf{v} \in \mathbb{R}^d$ be any arbitrary fixed (deterministic) vector. Let $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ be an arbitrary random vector. Consider the function

$$g(\mathbf{x}) = \mathbf{v}^\top \mathbf{x}.$$

Because $g(\mathbf{x})$ is a function of multiple random variables, it is itself a random variable. Prove that the expectation of $g(\mathbf{x})$ is given by:

$$\mathbb{E}[g(\mathbf{x})] = \mathbf{v}^\top \mathbb{E}[\mathbf{x}].$$

The above proof should be one line. Also, prove that

$$\mathrm{Var}(g(\mathbf{x})) = \mathbf{v}^\top \Sigma \mathbf{v},$$

where $\mathrm{Var}(\mathbf{x}) = \Sigma \in \mathbb{R}^d$. State why this implies that Σ is positive semidefinite.

The property in Problem 1(e) tells us that the covariance matrix Σ essentially encodes the variance of any *linear combination* of the entries of the random vector $\mathbf{x} = (x_1, \ldots, x_d)$. The coefficients of this linear combination are given by $\mathbf{v} = (v_1, \ldots, v_d)$.. Let's think about bacon, egg, and cheese sandwiches one final time to solidify this intuition.

Suppose your favorite corner breakfast cart is mathematically inclined and wants to understand the price fluctuations of two signature sandwiches: the *bacon egg and cheese* (BEC) and the *egg and cheese* (EC). Consider the 3-dimensional random vector $\mathbf{x} = (x_1, x_2, x_3)$ modeling the price of bacon (x_1), egg (x_2), and cheese (x_3), respectively. Suppose that the covariance matrix of this random vector is given by

$$\mathrm{Var}(\mathbf{x}) = \Sigma = \begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \end{bmatrix}$$

The recipe for BEC is 1 unit bacon, 1 unit egg, and 1 unit cheese, while the recipe for EC is 0 units bacon, 2 units egg, and 1 unit cheese. We can represent these as the vectors:

$$\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}.$$

Problem 1(f) [3 points] Using Problem 1(e), compute the variance of the price of a bacon, egg, and cheese, given by $\mathbf{b}^\top \mathbf{x}$. Compute the variance of the price of an egg and cheese, given by $\mathbf{e}^\top \mathbf{x}$.

In Problem 1(f), though we see that the price of bacon alone, given by $\mathrm{Var}(x_1) = 1.5$, fluctuates more than egg and cheese alone, the prices for egg and cheese are heavily correlated (with correlation $\mathrm{Cov}(x_2, x_3) = 2$), while the price of bacon is uncorrelated with the other two ingredients. That's why, as you should see from your computation, the fluctuations in the price of EC are greater than that of the BEC.

One might wonder, however: how did the corner breakfast cart magically stumble upon the true covariance matrix Σ of $\mathbf{x}$? Indeed, as we know from class, the job of statistics is to estimate such unknown parameters from observed data.

Recall that, in our statistical recasting of the regression problem we've been looking at all class, we imagine that, for our data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$

$$\mathbf{X} = \begin{bmatrix} \uparrow & \cdots & \uparrow \\ \mathbf{x}_1 & \cdots & \mathbf{x}_d \\ \downarrow & \cdots & \downarrow \end{bmatrix} = \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ \vdots & \vdots & \vdots \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix}$$

is a matrix of random variables, all drawn from some unknown probability distribution. To clarify how this connects with the notions in this problem, suppose that $\mathbf{x} \in \mathbb{R}^d$ is a

d-dimensional random vector, distributed as some unknown distribution $\mathbb{P}_{\mathbf{x}}$ over its random entries $x_1, \ldots, x_d$. You may think of $\mathbf{x}$ as the "true" random vector generating the data that we observe. For example, if we are thinking of the statistics for basketball players, as in PS1, x_1 might be a random variable representing height, x_2 might be a random variable representing weight, and so on. Written like this, the random variable $\mathbf{x}$ is yet unrealized. Because $\mathbf{x}$ is a random vector, it has a covariance matrix, which we will denote Σ:

$$\Sigma = \mathrm{Var}(\mathbf{x}) = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top]$$

When we collect data, we observe realizations of this random variable, and those are the rows of our data matrix, $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top \in \mathbb{R}^{1\times d}$. Each realization $\mathbf{x}_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$ are themselves d-dimensional random vectors drawn from some sampling process. Recall from lecture that our usual estimator, given i.i.d. realizations $X_1, \ldots, X_d$, for the mean of a random variable X is the *sample mean*:

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

This could be easily generalized to random vectors:

$$\overline{\mathbf{x}}_n := \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

for estimating $\mathbb{E}[\mathbf{x}]$. We might also want to estimate Σ from our data, and we do this through the *sample covariance matrix*, which we denote $\hat{\Sigma}_n$:

$$\hat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}}_n)(\mathbf{x}_i - \overline{\mathbf{x}}_n)^\top.$$

This matrix is intimately connected with the matrix $\mathbf{X}^\top\mathbf{X}$ that played a key part in the OLS estimator, as Problem 1(f) shows.

Problem 1(g) [4 points] Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the data matrix obtained as above. Suppose that each of the columns is centered — that is, suppose $\overline{\mathbf{x}}_n = \mathbf{0}$ (note that we can always pre-process a data matrix to obey this by first computing $\overline{\mathbf{x}}_n$ then subtracting if off). Prove that

$$\hat{\Sigma}_n = \frac{1}{n}\mathbf{X}^\top\mathbf{X}.$$

This shows us that, when recast in a statistical framework, the $\mathbf{X}^\top\mathbf{X}$ matrix in the OLS estimator, $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, is a (scaled) estimator of the true covariance matrix of the underlying random variable distribution generating our data. The $\frac{1}{n}$ factor difference comes from the fact that we have focused on optimizing the *sum* of squared residuals $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ instead of the *average* of squared residuals $\frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ for mathematical convenience.

Problem 2

Statistical properties of ridge regression [25 points].

In this problem, we will investigate a couple of statistical properties of the ridge regression estimator, motivating why, in some scenarios, it is preferable to the standard OLS estimator.

Recall, first, our error model, now fully fledged with our statistical understanding:

$$\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon,$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ are our training data, $\mathbf{y} \in \mathbb{R}^n$ are our training labels, $\mathbf{w}^* \in \mathbb{R}^d$ is the true linear relationship between $\mathbf{X}$ and $\mathbf{y}$, and $\epsilon \in \mathbb{R}^n$ is a vector of independent, mean-zero errors ϵ_i. That is, $\mathbb{E}[\epsilon_i] = 0$ and ϵ_i is independent of ϵ_j for all $i \neq j$. Each ϵ_i is also independent of $\mathbf{X}$. Let $\hat{\mathbf{w}}_{\text{ols}} := (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$ denote the standard OLS estimator throughout this problem. We have already shown in class that conditional on $\mathbf{X}$,

$$\mathbb{E}[\hat{\mathbf{w}}_{\text{ols}} \mid \mathbf{X}] = \mathbf{w}^*.$$

We will now build up to analyzing the conditional variance of this estimator, $\text{Var}(\hat{\mathbf{w}}_{\text{ols}} \mid \mathbf{X})$.

Problem 2(a) [3 points] For the rest of this problem, assume the error model above. Prove that:

$$\hat{\mathbf{w}}_{\text{ols}} - \mathbb{E}[\hat{\mathbf{w}}_{\text{ols}} \mid \mathbf{X}] = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \epsilon.$$

Problem 2(a) is helpful in analyzing the conditional variance of $\hat{\mathbf{w}}_{\text{ols}}$. Recall that the conditional variance of a random variable X conditioned on a random variable Y can just be obtained by replacing expectation by conditional expectation in the definition of variance:

$$\text{Var}(X \mid Y) = \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2 \mid Y] = \mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2 \tag{1}$$

Similarly, the conditional variance of a random vector $\mathbf{x}$ conditional on a random vector $\mathbf{y}$ is the covariance matrix defined by:

$$\text{Var}(\mathbf{x} \mid \mathbf{y}) = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}])(\mathbf{x} - \mathbb{E}[\mathbf{x} \mid \mathbf{y}])^\top \mid \mathbf{y}].$$

If $\mathbf{y}$ is a random matrix $\mathbf{Y}$ instead, the definition simply follows from replacing $\mathbf{y}$ with $\mathbf{Y}$. Throughout this problem, we will need the following property of the random errors ϵ.

Problem 2(b) [3 points] Prove that

$$\mathbb{E}[\epsilon\epsilon^\top \mid \mathbf{X}] = \sigma^2\mathbf{I}.$$

Problem 2(b) is a key property that allows us to derive an expression for the variance, as follows.

Problem 2(c) [4 points] Prove that

$$\text{Var}(\hat{\mathbf{w}}_{\text{ols}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}.$$

Problem 2(a) and Problem 2(b) will be helpful in this.

To get the *unconditional* variance, which averages over all possible realizations of the training data $\mathbf{X}$, we need to use a property analogous to the law of total expectation. Recall that for a pair of random variables X and Y, the *law of total expectation* states that:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X].$$

This useful property allows us to simplify unconditional expectations by conditioning on random variables, averaging over all the conditioning events. This immediately implies that the unconditional expectation of $\hat{\mathbf{w}}_{\text{ols}}$ is:

$$\mathbb{E}[\hat{\mathbf{w}}_{\text{ols}}] = \mathbf{w}^*.$$

There is an analogous law for conditional variances, known as the *law of total variance*:

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X \mid Y)] + \text{Var}(\mathbb{E}[X \mid Y]).$$

Recall from lecture that $\text{Var}(X \mid Y)$ and $\mathbb{E}[X \mid Y]$ are both *random variables*, as conditional expectations of random variables are random variables. We will prove the law of total variance in the next couple of exercises. Let X and Y be two arbitrary random variables.

Problem 2(d) [4 points] Prove that

$$\mathbb{E}[X^2] = \mathbb{E}\left[\text{Var}(X \mid Y) + \mathbb{E}[X \mid Y]^2\right].$$

It may be helpful to use the law of total expectation to introduce Y and then use the second characterization in Equation (1).

Problem 2(e) [3 points] Using Problem 2(c), prove that

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[\text{Var}(X \mid Y)] + \mathbb{E}[\mathbb{E}[X \mid Y]^2] - \mathbb{E}[\mathbb{E}[X \mid Y]]^2.$$

Using this and the fact that $\mathbb{E}[X \mid Y]$ is itself a random variable, conclude the law of total variance:

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X \mid Y)] + \text{Var}(\mathbb{E}[X \mid Y]).$$

You can use, without proof, that this generalizes to random vectors and matrices. Using this fact, show that

$$\text{Var}(\hat{\mathbf{w}}_{\text{ols}}) = \sigma^2 \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1}].$$

This fills in the gaps we left in lecture about the statistical properties of the OLS estimator.

Now, we consider the bias and variance of the *ridge regression estimator*. Recall that, for a tunable hyperparameter $\gamma > 0$, the ridge regression estimator minimizes the objective function:

$$f(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \gamma\|\mathbf{w}\|^2.$$

We showed that the solution is given by:

$$\hat{\mathbf{w}}_{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}.$$

For the remaining problems, we will only consider the *conditional* expectation and variances, given $\mathbf{X}$, which should simplify calculations. One of the main reasons one might use the ridge estimator instead of the OLS estimator is that, for certain choices of γ, the ridge estimator incurs more bias to reduce the variance of the estimator. In particular, in many cases, this results in the ridge estimator having better *risk* than the OLS estimator. We will not do a formal analysis of the risk of the ridge estimator, but we will derive quantities for its bias and variance.

First, we will show that the ridge estimator is no longer unbiased, unlike the OLS estimator.

Problem 2(f) [4 points] Show that, conditional on $\mathbf{X}$, the "bias" of the $\hat{\mathbf{w}}_{\text{ridge}}$ in estimating $\mathbf{w}^*$ is given by:

$$\mathbb{E}[\hat{\mathbf{w}}_{\text{ridge}} \mid \mathbf{X}] - \mathbf{w}^* = -\gamma(\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-1}\mathbf{w}^*.$$

Hint: To get $\mathbb{E}[\hat{\mathbf{w}}_{\text{ridge}} \mid \mathbf{X}] - \mathbf{w}^*$, it may help to add and subtract $\gamma\mathbf{I}$ to an appropriate quantity.

We will not analyze this quantity further, but one can show that, for a fixed $\mathbf{x}_0$, when considering the squared bias of $\mathbf{x}_0^\top\hat{\mathbf{w}}_{\text{ridge}}$ to the estimand $\mathbf{x}_0^\top\mathbf{w}^*$, we obtain a quantity that grows with γ, aligning with our intuition that increasing the γ parameter in the ridge objective biases us towards "smaller" $\hat{\mathbf{w}}$ values. To heuristically see this, as an optional exercise, compute the squared norm of the "bias" you found in Problem 2(e).

Finally, we'll find an expression for the variance of $\hat{\mathbf{w}}_{\text{ridge}}$, the covariance matrix.

Problem 2(g) [4 points] Prove that, for any $\gamma > 0$, the matrices $\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I}$ and $\mathbf{X}^\top\mathbf{X}$ commute. That is,

$$(\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})\mathbf{X}^\top\mathbf{X} = \mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I}).$$

Using this property and Problem 2(b), prove that

$$\text{Var}(\hat{\mathbf{w}}_{\text{ridge}} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-2}.$$

It may also help to use the facts that $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$ and $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$

We will not explore this quantity further, but recall from lecture that, performing an eigen-decomposition of the matrix $(\mathbf{X}^\top\mathbf{X} + \gamma\mathbf{I})^{-2} = \mathbf{V}(\mathbf{\Lambda} + \gamma\mathbf{I})^{-2}\mathbf{V}^\top$ where the eigenvalues are given by $\frac{1}{(\lambda_i+\gamma)^2}$, which are always bounded

$$\frac{1}{(\lambda_i + \gamma)^2} \leq \frac{1}{\gamma^2}.$$

Problem 3

Course evaluations (0 points total, but very much appreciated by me).

Congratulations on getting to the end of the course! I hope you've learned a thing or two. There's no programming component to this problem set, so all you have to turn in is your PDF file.

Optional, but very much appreciated: If you could take a minute to fill out the official SEAS course evaluations on Courseworks, that'd be really appreciated. As you know, this is the first iteration of this course, and I'd really appreciate it if I could get feedback as an instructor as well as feedback on what went well or poorly.

Also optional, but very much appreciated: As a more specific evaluation of the course, it'd also be really nice if you could fill out the anonymous post-course survey.

Very optional: If you plan on taking *COMS 4771: Machine Learning* next semester or in a future semester, I'd love to just hear about how you're doing in the course, if there's any mathematical preliminary in 4771 that you feel should've been covered more in my class, and just how you feel about the course (or future ML courses) in general.

Very, very optional: Or, if you plan on doing ML research in the future, I'd love to chat about your plans!