Solutions

Homework 0: review of probability, statistics, linear algebra, and calculus

CSCI145/MATH166, Mike Izbicki

Problem 1. (10 pts) A couple people missed points on this problem for not defining the terms.

Problem 2. (15 pts) Prove or give a counter example: If two random variables x,y are independent, then their covariance is zero.

Proof. First, we show that

$$\begin{aligned}
\mathbb{E}_{x,y}xy &= \iint xy\ p(x,y)\ dx\ dy && (1)\\
&= \iint xy\ p(x)\ p(y)\ dx\ dy && (2)\\
&= \left(\int x\ p(x)\ dx\right)\left(\int y\ p(y)\ dy\right) && (3)\\
&= (\mathbb{E}_x x)(\mathbb{E}_y y) && (4)
\end{aligned}$$

Then by the definition of covariance,

$$\mathrm{Cov}(x,y) = \mathbb{E}_{x,y}xy - (\mathbb{E}_x x)(\mathbb{E}_y y) = 0. \tag{5}$$

□

Grading notes: I was originally expecting you to explicitly prove equation (4), but decided to give you full credit even if you just assumed it.

Problem 3. (15 pts) Prove or give a counter example: If two random variables x,y have zero covariance, then they are independent.

Example wrong (but close) answer: This statement is false. As a counterexample, let x be a random variable with $\mathbb{E}x = 0$ and $\mathbb{E}x^3 = 0$, and let $y = x^2$. Then x and y are not independent, but

$$\begin{aligned}
\mathrm{Cov}(x,y) &= \mathbb{E}xy - \mathbb{E}x\mathbb{E}y && (6)\\
&= \mathbb{E}x^3 - \mathbb{E}x\mathbb{E}y && (7)\\
&= 0. && (8)
\end{aligned}$$

Why is this wrong? The definition of x and y do not imply that x and y are dependent. For example, let x be the random variable that takes on values of 1 or -1 each with probability 1/2. Then x satisfies the assumptions above (i.e. $\mathbb{E}x = 0$ and $\mathbb{E}x^3 = 0$). No matter what the value of x is, y will always be equal to 1. And so in this case, x and y are independent. An even simpler example is if x and y are both the constant zero random variables.

How to fix it? Whenever you provide a counterexample, it is not enough to describe the counterexample. You must explicitly state what the example is. In this case, if we let x be the uniform distribution over $[-1, 1]$, then x and y are actually dependent on each other, and the remainder of the argument holds.

Problem 4. (10pts) Everyone got this mostly correct.

Problem 5. (5 pts) Everyone got full credit.

Problem 6. (5 pts) Everyone got full credit.

Problem 7. (10 pts) Everyone got full credit.

Problem 8. (15 pts) The log-loss is defined to be

$$\ell(\mathbf{x}; \mathbf{w}) = \log(1 + \exp(-\mathbf{x}^T \mathbf{w})). \tag{9}$$

Calculate the first and second derivatives of (9) with respect to $\mathbf{w}$. Note that because $\mathbf{w}$ is a vector, the loss is a scalar, the derivative is a vector, and the second derivative is a matrix. See Section 2 of *The Matrix Cookbook* for a review of vector derivatives.

The first derivative is:

$$\begin{aligned}
\frac{d}{d\mathbf{w}} \ell(\mathbf{x}; \mathbf{w}) &= \frac{d}{d\mathbf{w}} \log(1 + \exp(-\mathbf{x}^T \mathbf{w})) && (10)\\
&= \frac{1}{1 + \exp(-\mathbf{x}^T \mathbf{w})} \frac{d}{d\mathbf{w}} (1 + \exp(-\mathbf{x}^T \mathbf{w})) && (11)\\
&= \frac{-\mathbf{x} \exp(-\mathbf{x}^T \mathbf{w})}{1 + \exp(-\mathbf{x}^T \mathbf{w})} && (12)\\
&= \frac{-\mathbf{x}}{1 + \exp(\mathbf{x}^T \mathbf{w})} && (13)
\end{aligned}$$

The second derivative is:

$$\begin{aligned}
\frac{d}{d\mathbf{w}} \frac{d}{d\mathbf{w}} \ell(\mathbf{x}; \mathbf{w}) &= \frac{d}{d\mathbf{w}} - \mathbf{x}(1 + \exp(\mathbf{x}^T \mathbf{w})))^{-1} && (14)\\
&= \frac{d}{d\mathbf{w}} - \mathbf{x}(1 + \exp(\mathbf{x}^T \mathbf{w})))^{-1} && (15)\\
&= \left[\frac{d}{d\mathbf{w}} (1 + \exp(\mathbf{x}^T \mathbf{w}))\right] \left[\mathbf{x}^T (1 + \exp(\mathbf{x}^T \mathbf{w})))^{-2}\right] && (16)\\
&= \left[\mathbf{x}(\exp(\mathbf{x}^T \mathbf{w}))\right] \left[\mathbf{x}^T (1 + \exp(\mathbf{x}^T \mathbf{w})))^{-2}\right] && (17)\\
&= \mathbf{x}\mathbf{x}^T (\exp(\mathbf{x}^T \mathbf{w}))(1 + \exp(\mathbf{x}^T \mathbf{w})))^{-2} && (18)
\end{aligned}$$

Grading notes:

1. if you had the type of the derivatives being scalars (instead of vectors/matrices), you lost 5 pts each
2. You cannot square a vector, i.e. x^2 does not make sense. Similarly, you cannot multiply a vector times itself; $\mathbf{x}\mathbf{x}$ and $\mathbf{x}^T \mathbf{x}^T$ don't make sense. -3 for this mistake.

Problem 9. (15 pts) Calculate

$$\arg\min_{\mathbf{w}} \left(\|\mathbf{y} - X\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|^2 \right) \tag{19}$$

where $\mathbf{w}$ is a vector of dimension d, $\mathbf{y}$ is a vector of dimension n, and X is a matrix with shape $n \times d$.

HINT: Recall that the arg min is defined as

$$\arg\min_{\mathbf{w}} f(\mathbf{w}) \triangleq \{\mathbf{w} : f(\mathbf{w}) = \min_{\mathbf{w}'} f(\mathbf{w}')\}. \tag{20}$$

That is, the arg min returns the set of values that minimize the function f. To calculate the arg min, take the derivative of f with respect to $\mathbf{w}$, set it equal to zero, and solve for $\mathbf{w}$.

$$\begin{align}
0 &= \frac{d}{d\mathbf{w}} \left[\|\mathbf{y} - X\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|^2 \right] \tag{21} \\
&= \frac{d}{d\mathbf{w}} \|\mathbf{y} - X\mathbf{w}\|^2 + \frac{d}{d\mathbf{w}} \lambda\|\mathbf{w}\|^2 \tag{22} \\
&= -2X^T(\mathbf{y} - X\mathbf{w}) + 2\lambda\mathbf{w} \tag{23} \\
&= -X^T\mathbf{y} + X^TX\mathbf{w} + \lambda\mathbf{w} \tag{24} \\
X^T\mathbf{y} &= X^TX\mathbf{w} + \lambda\mathbf{w} \tag{25} \\
&= (X^TX + \lambda I)\mathbf{w} \tag{26} \\
(X^TX + \lambda I)^{-1}X^T\mathbf{y} &= \mathbf{w} \tag{27}
\end{align}$$

Grading notes:

1. Let A be a square matrix and λ a scalar. $A + \lambda$ is different than $A + \lambda I$. The former adds the constant to every element, the latter only along the diagonal. -2 points for leaving out the I.
2. You cannot divide by a matrix, you must multiply by the inverse. This is because matrix multiplication is not commutative, so you must specify whether it's a left or right multiplication. -2 for this mistake.

Problem 10. (2pts extra credit) If you complete this assignment in LaTeX, then you will receive +2 pts extra credit. The source files are available at:
`https://github.com/mikeizbicki/cmc-csci145-math166/tree/master/hw_00`.

Grading notes:

1. In the future, you will need to have better typesetting to get the LaTeX extra credit.
2. You cannot put text directly in math mode. For example,

   $$log(1 + exp(a)) \tag{28}$$

   is bad. Instead, do

   $$\log(1 + \exp(a)) \tag{29}$$

   by typing `\log` and `\exp` instead of just `log` and `exp`.
3. Similarly, you should use $\mathbb{E}$ not E, max not max, etc.
4. We do not use `*` for multiplication ever. Only juxtaposition or $\cdot$ (`\cdot`).
5. If you have a multiline equation, then you must use the `align` environment and have the $=$ properly aligned.
6. I don't care whether you use x or X type variables, but you must be consistent.