Quiz 2 Solutions

02 Ridge Regression

by Alvin Wan . alvinwan.com/cs189/fa17

Treat this as an exam situation. You will be given 5 minutes to complete this quiz.

1 Probability Review

1. If $y = Ax + b$ and $\text{cov}(x) = \Sigma$ (A, b are knowns), compute $\text{cov}(y)$

   Solution:

   First, notice $y - E[y] = (Ax+b) - E[Ax+b] = Ax+b-(AE[x]+b) = A[x-E[x]]$.

$$\begin{aligned}\text{cov}(y) &= E[(y - E[y])(y - E[y])^T] \\ &= AE[(x - E[x])(x - E[x])^T]A^T \\ &= A\Sigma A^T\end{aligned}$$

2. (more linear algebra-ish) Prove that if X is a valid transition matrix (i.e., the entries of every row sum to 1), then X^k for any positive integer k will yield a matrix with the same property.

   Solution: Note that if the matrix A is a valid transition matrix, $A\mathbb{1} = \mathbb{1}$ for the vector of all ones $\mathbb{1}$.

   Base case: $XX\mathbb{1} = X\mathbb{1} = \mathbb{1}$

   Inductive Hypothesis: Assume this holds for X^i.

   Inductive Step: $X^{i+1}\mathbb{1} = X^iX\mathbb{1} = X^i\mathbb{1} = \mathbb{1}$, where we apply the inductive hypothesis in the last step.

2 Generalized Tikhonov Regularization

Here we explore a generalized version of ridge regression, Tikhonov regularization:

$$\min_w \|Xw - y\|_2^2 + \|\Gamma w\|_2^2$$

1. We can even generalize Tikhonov regularization, if we view w as general multivariate gaussians. Find a closed-form solution to the following optimization problem, where $\mu = E[w]$, $\|x\|_A^2 = x^T Ax$ and A, B are inverse covariance matrices of y, w, respectively.

$$\min_w \|Xw - y\|_A^2 + \|w - \mu\|_B^2$$

Solution:

Keep in mind that we've previously proved $\frac{\partial x^T w}{\partial x} = \frac{\partial w^T x}{\partial x} = w$ and $\frac{\partial x^T Ax}{\partial x} = (A + A^T)x$ (see crib 1).

$$\begin{aligned}
&\frac{\partial}{\partial w}(Xw - y)^T A(Xw - y) + (w - \mu)^T B(w - \mu) \\
&\overset{(a)}{=} \frac{\partial}{\partial w}(w^T X^T AXw - 2w^T X^T Ay + y^T y + w^T Bw - 2w^T B\mu + \mu^T \mu) \\
&= X^T(A + A^T)Xw - 2X^T Ay + (B + B^T)w - 2B\mu \\
&= 2X^T AXw - 2X^T Ay + 2Bw - 2B\mu \\
&= 2X^T A(Xw - y) + 2B(w - \mu)
\end{aligned}$$

(a) Remember: we can mush $a^T b = b^T a$ for column vectors a, b. (b) Covariance matrices are symmetric, so $B = B^T$. Now, set equal to 0 and solve.

$$\begin{aligned}
2X^T A(Xw - y) + 2B(w - \mu) &= 0 \\
X^T A(Xw - y) &= B(\mu - w) \\
(X^T AX + B)w &= X^T Ay + B\mu \\
w^* &= (X^T AX + B)^{-1}(X^T Ay + B\mu)
\end{aligned}$$

2. Relate this generalized expression to ridge regression as formulated in class:

$$\min_w \|Xw - y\|_2^2 + \lambda\|w\|_2^2$$

Solution:

We make the assumption that all y are i.i.d. Then, the inverse covariance matrix for y, $A = I$, is the identity. Note that using our definition from the previous part of $\|x\|_A^2 = x^T Ax = x^T x = \|x\|_2^2$. Thus,

$$\|Xw - y\|_A^2 = \|Xw - y\|_2^2$$

Pick a suitable value for $B = \Gamma^T\Gamma$, for a Tikhonov matrix Γ. We use $\Gamma = \lambda I$, meaning we assume all w_i are i.i.d. with variance λ. Thus, $\|w\|_B^2 = \lambda w^T w = \lambda\|w\|_2^2$

$$\|w - \mu\|_B^2 = \lambda\|w - \mu\|_2^2$$

Finally, we assume w is 0-mean, so $\mu = 0$. This gives us our final formulation of ridge regression.

$$\min_w \|Xw - y\|_2^2 + \lambda\|w\|_2^2$$

In short, ridge regression assumes that w are i.i.d. 0-mean with variance λ and that y are i.i.d.

3. Using your understanding of the relationship, find the closed-form solution to ridge regression. (i.e., You should not need gradients.)

   Solution: From the previous part, we know this specific case sets $A = I, B = \lambda I, \mu = 0$, so plugging in, we have:

$$w^* = (X^T AX + B)^{-1}(X^T Ay + B\mu) = (X^T X + \lambda I)^{-1} X^T y$$