ICS635 Homework 1:

by: Lambert Leong

Problem 1

1.

$$\begin{aligned}
E[aX+b] &= \int_{-\infty}^{\infty}(aX+b)f(x)dx \\
&= \int_{-\infty}^{\infty} aXf(x)dx + \int_{-\infty}^{\infty} bf(x)dx \\
&= a\int_{-\infty}^{\infty} Xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx \\
&= aE[X]+b
\end{aligned}$$

2.

$$\begin{aligned}
var(cX) &= E[(cX-c\mu)^2] \\
&= E[cX]^2 - (E[cX])^2 \\
&= c^2E[X^2] - c^2(E[X])^2 \\
&= c^2(E[X^2] - (E[X])^2) \\
&= c^2(E[X] - (E[X])^2) \\
&= c^2 var(X)
\end{aligned}$$

3.

$$\begin{aligned}
var(x) &= \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx \\
&= \int_{-\infty}^{\infty}(x^2+\mu^2-2x\mu)f(x)dx \\
&= \int_{-\infty}^{\infty} x^2 f(x)f(x) + \mu^2\int_{-\infty}^{\infty} f(x)dx - 2\mu\int_{-\infty}^{\infty} xf(x)dx \\
&= E[X^2]+\mu^2-2\mu^2 \\
&= E[X^2]-\mu^2 \\
&= E[X^2]-E[X]^2
\end{aligned}$$

Problem 2

1.

$$\begin{aligned}
E[X] &= \int_a^b x(\tfrac{1}{b-a})dx \\
&= \tfrac{b+a}{2}
\end{aligned}$$

2.

$$\begin{aligned} var(X) &= E[X^2] - E^2[X] \\ &= \int_a^b \tfrac{x^2}{b-a} dx - (\tfrac{b-a}{2})^2 \\ &= \frac{b^2 + ba + a^2}{3} - (\tfrac{b+a}{2})^2 \\ &= \frac{b^2 - 2ba + a^2}{12} \end{aligned}$$

Problem 3

1.

$$F(x) = P(X \leq x) = 1 - P(X > x) = 1 - P(min(X_1, X_2) > x)$$

If X_1 and X_2 are evenly distributed on $[0, 1]$

$$\begin{aligned} F(x) &= 1 - \left(\frac{b-x}{b-a}\right)^n = 1 - \left(\frac{1-x}{1-0}\right)^2 = 1 - (1-x)^2 \\ f(x) &= \frac{dF}{dx} = 2\,(-x+1) \\ E[X] &= \int_0^1 xf(x)dx = \int_0^1 x2(-x+1)dx = \frac{1}{3} \end{aligned}$$

2.

$$\begin{aligned} var(X) &= E[X^2] - E^2[X] \\ &= \int_0^1 x^2 2(-x^2+1)dx - \left(\int_0^1 x2(-x+1)dx\right)^2 \\ &= \frac{7}{45} \end{aligned}$$

3.

$$cov(X, X_1) = E[XX_1] - E[X]E[X_1]$$

Problem 4

1.

$$P(D|T^+) = \tfrac{P(T^+|D)P(D)}{P(T^+)}$$

2.

$$\begin{aligned} P(D|T^+) &= \tfrac{P(T^+|D)P(D)}{P(T^+)} \\ &= \tfrac{P(T^+|D)P(D)}{[P(T^+|D)\times P(D)]+[P(T^+|notD)\times P(notD)]} \\ &= \tfrac{SP(D)}{[SP(D)]+[(1-Q)(1-P(D))]} \\ &= \tfrac{.99(.001)}{[.99(.001)]+[(1-.99)(1-.001)]} \\ &= .0902 \end{aligned}$$

3.

$$\begin{aligned} P(D|T^+) &= \tfrac{P(T^+|D)P(D)}{P(T^+)} \\ &= \tfrac{P(T^+|D)P(D)}{[P(T^+|D)\times P(D)]+[P(T^+|notD)\times P(notD)]} \\ &= \tfrac{SP(D)}{[SP(D)]+[(1-Q)(1-P(D))]} \\ &= \tfrac{.99(.001)}{[.99(.001)]+[(1-.90)(1-.001)]} \\ &= .0098 \end{aligned}$$

Problem 5

1.

$$\nabla f(x) = \tfrac{1}{2}[\nabla x^T A x] + b^T$$

if A symmetric

$$\vec{\nabla}(x^T A x) = Ax + x^T A = 2Ax$$

Then

$$\begin{aligned} &= \tfrac{1}{2}(2Ax) + b^T x \\ &= Ax + b^T \end{aligned}$$

2.

$$\begin{aligned} \nabla f(x) &= g(h(x)) \\ &= g'(h(x))h'(x) \end{aligned}$$

3.

$$\nabla^2 f(x) = \frac{1}{2}(x^T A x) = Ax + x^T A = 2Ax = A$$

4.

$$\begin{aligned}\nabla f(x) &= \nabla g(a^T x)\\ &= g'(a^T x)\nabla a^T x\\ &= (g'(a^T x))a\end{aligned}$$

$$\begin{aligned}\nabla^2 f(x) &= \nabla(g'(a^T x))a\\ &= a^T a g''(a^T x)\end{aligned}$$

Problem 6

1. (a) A small λ would cause overfiting of the on the training set and lead to low error.
   (b) The overfit model will not generalize to the validation set well and cause more error
   (c) w will decrease
   (d) The number of non-zero elemnts in w would be lower
2. (a) A big λ could lead to under fitting on the training set which would increase error
   (b) Underrfitting on the training set could lead to better generalization of the validation set and reduce error.
   (c) w magnitude will increase.
   (d) There will be more non-zero elements of w.
3. (a) An L2 regularization encourages non-zero elements towards zero but they dont actuall become zero.
   (b)

$$\begin{aligned}\nabla L &= \sum_{n=1}^{N} \tfrac{\partial L}{\partial w}(y_n - (w^T x_n))^2\\ &= \sum_{n=1}^{N} -2x_n(y_n - w^T x_n)\end{aligned}$$

   (c)

$$\begin{aligned}\nabla_\theta \mathcal{L}(\mathcal{D};\theta,\lambda) &= \sum_{n=1}^{N} \tfrac{\partial \mathcal{L}}{\partial w}(y_n - (w^T x_n))^2 + \lambda||w||_2\\ &= \sum_{n=1}^{N} \tfrac{\partial \mathcal{L}}{\partial w}(y_n - (w^T x_n))^2 + \lambda\sum_{d=1}^{D} \tfrac{\partial \mathcal{L}}{\partial w} w_d^2\\ &= \sum_{n=1}^{N} -2x_n(y - w^T x_n) + \lambda\sum_{d=1}^{D} 2w_d\end{aligned}$$

Problem 7

1. The outcomes of the XOR functions are not linearly sperable. Therefore, no parameters exists that can lead to a zero loss for a single neuron.

2.

$$\begin{aligned}
w_1^{(1)} &= 1, w_2^{(1)} = 1, b^1 = -.5 \\
w_1^{(2)} &= -1, w_1^{(2)} = -1, b^2 = 1.5 \\
w_1^{(3)} &= 1, w_2^{(3)} = 1, b^3 = -1.5
\end{aligned}$$

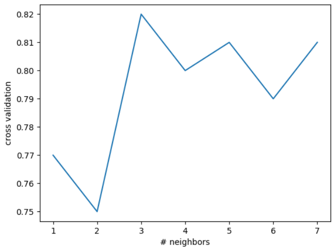
$$\begin{aligned}
h_1 &= (1 * x_1 + 1 * x_2 + (-0.5) \\
h_2 &= (-1) * x_1 + (-1) * x_2 + (1.5) \\
y &= (1 * (h_1) + (1) * (h_2) + (-1.5)
\end{aligned}$$

Problem 8

1. Feature scaling for linear classifiers will have little not no effect due to the fact that weights are added to features and often times a bias term exists. K-nearest neighbors (k-nn), on the other hand, measures the euclidean distance and will be more affected by scaling. Scaling for k-nn would help to prevent a model from being dominated by a feature with a large range.

2. The best accuracy for training and validation is 0.88 and 0.92 respectively. The income histogram indicated a lot of overlap between the two classes. Also, class 0 seemed to have some outliers. Outliers were determined by x amount of standard deviations (std) away from the mean. While it is common to consider data points outside of two std's from the mean to be an outlier, better training and validation accuracies were seen when we excluded points that were outside of 1 std.

   We used a K-nearest neighbors (Knn) model and experiemented with different hyperparameters such as the algorithm used to compute the nearest neighbors, weight function used in prediction, and the number of neighbors. Ultimately, the number of neighbors seemed to have the greatest effect on the accuracy.



3. The number of neighbors is the hyperparameter I chose to optimize. It appears that it is best to used an odd number of neighbors. Three neighbors lead to the best cross validation accuracy. Any more and any less seem to yeild lower accuracies.