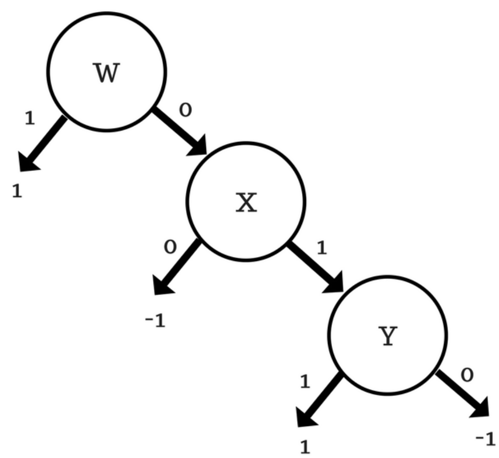
Midterm Practice Problems Solutions

Problem 1 (Decision Tree)

(a) $H(Z) = H(\frac{2}{5}) = \frac{2}{5}\log(\frac{5}{2}) + \frac{3}{5}\log(\frac{5}{3}) = 0.970951$

$$\begin{aligned}
H(Z|split(W)) &= \frac{1}{5}H(Z|branch(W=1)) + \frac{4}{5}H(Z|branch(W=0)) \\
&= \frac{1}{5}H(1) + \frac{4}{5}H(\frac{1}{4}) = 0.649 \\
H(Z|split(X)) &= \frac{2}{5}H(Z|branch(X=1)) + \frac{3}{5}H(Z|branch(X=0)) \\
&= \frac{2}{5}H(0.5) + \frac{3}{5}H(\frac{1}{3}) = 0.951 \\
H(Z|split(Y)) &= \frac{2}{5}H(Z|branch(Y=1)) + \frac{3}{5}H(Z|branch(Y=0)) \\
&= \frac{2}{5}H(0.5) + \frac{3}{5}H(\frac{1}{3}) = 0.951
\end{aligned}$$

Therefore splitting with W has the highest information gain of 0.321951 in comparison to splitting with X or Y both having IG of 0.0199



(b)

X and Y can be swapped to get another possible decision tree.

(c) Samples# 5 and 7 have the same features but different labels, therefore it is not possible to construct such a decision tree.

Problem 2 (Perceptron)

(a) AND

$\theta = (2, 2, -3)$ if the augmented features are $(x_1, x_2, 1)$. Multiple solutions are possible.

(b) XOR

No solution exists because the data is not linearly separable.

Problem 3 (Logistic Regression)

(a) For finding $\nabla_w J(\boldsymbol{w}, \boldsymbol{x})$ first note that $\alpha_i(\boldsymbol{x})$ doesn't depend on $\boldsymbol{w}$. Therefore the derivation will be similar to that for logistic regression except for the additional localized weights. $\nabla_w J(\boldsymbol{w}, \boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i(\boldsymbol{x})(h_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i)\boldsymbol{x}_i$.
Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, then we can write $\nabla_w J(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{X^T z}$, where $\boldsymbol{z}$ is a vector of $\mathbb{R}^n$ with each entree $\alpha_i(\boldsymbol{x})(h_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i)$.

(b) From the expression for the gradient you can see that

$$\frac{\partial J(\boldsymbol{w}, \boldsymbol{x})}{\partial w_j} = \sum_{i=1}^{n} \alpha_i(\boldsymbol{x})(h_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i)x_{i,j}$$

$$\begin{aligned}
\frac{\partial^2 J(\boldsymbol{w}, \boldsymbol{x})}{\partial w_k \partial w_j} &= \frac{\partial}{\partial w_k}\Big(\sum_{i=1}^{n} \alpha_i(\boldsymbol{x})(h_{\boldsymbol{w}}(\boldsymbol{x}_i) - y_i)x_{i,j}\Big) \\
&= \sum_{i=1}^{n} \alpha_i(\boldsymbol{x})\frac{\partial}{\partial w_k} h_{\boldsymbol{w}}(\boldsymbol{x}_i)x_{i,j} \\
&= \sum_{i=1}^{n} \alpha_i(\boldsymbol{x})h_{\boldsymbol{w}}(\boldsymbol{x}_i)(1 - h_{\boldsymbol{w}}(\boldsymbol{x}_i))x_{i,j}x_{i,k}
\end{aligned}$$

Therefore we have

$$\nabla_w^2 J(\boldsymbol{w}, \boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i(\boldsymbol{x})h_{\boldsymbol{w}}(\boldsymbol{x}_i)(1 - h_{\boldsymbol{w}}(\boldsymbol{x}_i))\boldsymbol{x}_i\boldsymbol{x}_i^T = \boldsymbol{X^T D X}$$

$\boldsymbol{u^T X^T D X u} = ||D^{\frac{1}{2}}Xu||_2^2 > 0 \quad \forall \boldsymbol{u} \neq 0$

(c)

$$\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t - \eta X^T z^t$$

(d) Non parametric method.

Problem 4 (Linear Regression)

(a)

(b) Gradient Descent:

$$w_1^{t+1} \quad \leftarrow \quad w_1^t + \eta \sum_{i=1}^{M} (y_i - (w_1^t x_1^{(i)} + w_2^t x_2^{(i)})) x_1^{(i)} \tag{1}$$

$$w_2^{t+1} \quad \leftarrow \quad w_2^t + \eta \sum_{i=1}^{M} (y_i - (w_1^t x_1^{(i)} + w_2^t x_2^{(i)})) x_2^{(i)} \tag{2}$$

You can also do stochastic gradient descent.

(c) To prove Eq. (**??**) has a global optimum, we have to show the function is convex. To prove the function is convex, we need to demonstrate the Hessian matrix (or the second derivatives) is positive semi-definite.

The Hessian of Eq. (**??**) is

$$H = \begin{bmatrix} \sum (x_1^{(i)})^2 & \sum x_1^{(i)} x_2^{(i)} \\ \sum x_1^{(i)} x_2^{(i)} & \sum (x_2^{(i)})^2 \end{bmatrix} \tag{3}$$

To show H is positive semi-definite, we have to prove for every vector $z \neq 0$, $z^T H z \geq 0$. This can be done by the following equations:

$$\begin{aligned} z^T H z &= \sum (x_1^{(i)})^2 z_1^2 + 2 \sum x_1^{(i)} x_2^{(i)} z_1 z_2 z_1 z_2 + \sum (x_1^{(i)})^2 z_2^2 \\ &= \sum (z_1 x_1^{(i)} + z_2 x_2^{(i)})^2 \\ &\geq 0 \end{aligned} \tag{4}$$

Problem 5 (Maximum Likelihood Estimation)

(a)

$$\begin{aligned} \mathcal{L}(x_1, x_2, \ldots, x_n; \lambda) &= \prod_{i=1}^{n} f(x_i | \lambda) \\ &= \lambda^n (x_1 x_2 \ldots x_n)^{-\lambda - 1} \end{aligned}$$

(b)

$$\begin{aligned}
arg\max_{\lambda} \mathcal{L}(x_1, x_2, \ldots, x_n; \lambda) &= arg\min_{\lambda} \Big(- \ln \mathcal{L}(x_1, x_2, \ldots, x_n; \lambda) \Big) \\
-\mathcal{L}\mathcal{L} &= -n \ln \lambda + (\lambda + 1) \sum_{i=1}^{n} \ln x_i \\
\frac{d(-\mathcal{L}\mathcal{L})}{d\lambda} &= -\frac{n}{\lambda} + \sum_{i=1}^{n} \ln x_i = 0 \\
\lambda_{mle} &= \frac{n}{\sum_{i=1}^{n} \ln x_i}
\end{aligned}$$

But since $\lambda > 1$ is known about the model, $\lambda_{mle} = \max(1, \frac{n}{\sum_{i=1}^{n} \ln x_i})$

(c) $\lambda_{mle} = \max(1, \frac{4}{6}) = 1$

Problem 6 (Maximum Likelihood Estimation 2)

(a)

$$\begin{aligned}
l(\theta) &= \sum_{n} \log P(X_i; \theta) \\
&= \sum_{n} x_n \log(\theta) + (1 - x_n) \log(1 - \theta) \\
&= 3 \log(\theta) + \log(1 - \theta)
\end{aligned}$$

(b)

$$l^{'}(\theta) = \frac{3}{\theta} - \frac{1}{1 - \theta}$$

(c)

$$\begin{aligned}
l^{'}(\theta) &= \frac{3}{\theta} - \frac{1}{1 - \theta} = 0 \\
\hat{\theta} &= \frac{3}{4}
\end{aligned}$$

Problem 7 (Kernel)

To show that K_β is a kernel, simply use the same feature mapping as used by the polynomial kernel of degree 3, but first scale $\mathbf{x}$ by $\sqrt{\beta}$. So, in effect they are both polynomial kernels of degree 3. If you look at the resulting feature vector, the offset term 1 is unchanged, the linear terms are scaled by $\sqrt{\beta}$, the quadratic terms are scaled by β, and the cubic terms are scaled by $\beta^{1.5}$. Although the model class remains unchanged, this changes how we penalize the features during learning (from the $||\theta||^2$ in the objective). In particular, higher-order features will become more costly to use, so this will bias more towards a lower-order polynomial.

That is,

$$\begin{aligned} K_\beta(\mathbf{x}, \mathbf{z}) &= (1 + \beta \mathbf{x} \cdot \mathbf{z})^3 \\ &= (1 + \beta (x_1 z_1 + x_2 z_2))^3 \\ &= 1 + 3\beta (x_1 z_1 + x_2 z_2) + 3\beta^2 \left(x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \right) \\ &\qquad + \beta^3 \left(x_1^3 z_1^3 + 3x_1^2 z_1^2 x_2 z_2 + 3x_1 z_1 x_2^2 z_2^2 + x_2^3 z_2^3 \right) \end{aligned}$$

so that

$$\phi_\beta(\mathbf{x}) = (1, \sqrt{3\beta} x_1, \sqrt{3\beta} x_2, \sqrt{3}\beta x_1^2, \sqrt{6}\beta x_1 x_2, \sqrt{3}\beta x_2^2, \sqrt{\beta^3} x_1^3, \sqrt{3\beta^3} x_1^2 x_2, \sqrt{3\beta^3} x_1 x_2^2, \sqrt{\beta^3} x_2^3)^T$$

The m^{th}-order terms in $\phi_\beta(\cdot)$ are scaled by $\beta^{m/2}$, so β trades off the influence of the higher-order versus lower-order terms in the polynomial. If $\beta = 1$, then $\beta^{1/2} = \beta = \beta^{3/2}$ so that $K_\beta = K$. If $0 < \beta < 1$, then $\beta^{1/2} > \beta > \beta^{3/2}$ so that lower-order terms have more weight and higher-order terms less weight; as $\beta \to 0$, K_β approaches $1 + 3\beta \mathbf{x} \cdot \mathbf{z}$ (a linear separator). If $\beta > 1$, the trade-off is reversed; as $\beta \to \infty$, only the constant and cubic terms in K_β remain.