Solution 7 - Optimization, LDA, and Naive Bayes

1. (a) We have

$$\begin{aligned} f(\mathbf{x}) = ||\mathbf{A}\mathbf{x} - \mathbf{b}||^2 &= (\mathbf{A}\mathbf{x} - \mathbf{b})^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= (\mathbf{x}^\top\mathbf{A}^\top - \mathbf{b}^\top)(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= \mathbf{x}^\top(\mathbf{A}^\top\mathbf{A})\mathbf{x} - 2(\mathbf{b}^\top\mathbf{A})\mathbf{x} + \mathbf{b}^\top\mathbf{b} \end{aligned}$$

which is a quadratic function. The gradient is given by $\nabla f(x) = 2(\mathbf{A}^\top\mathbf{A})\mathbf{x} - 2(\mathbf{A}^\top\mathbf{b})$, and the Hessian is given by $F(\mathbf{x}) = 2(\mathbf{A}^\top\mathbf{A})$.

(b) The fixed step size gradient algorithm for solving the above optimization problem is given by

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} - \alpha\left(2(\mathbf{A}^\top\mathbf{A})\mathbf{x}^{(k)} - 2\mathbf{A}^\top\mathbf{b}\right) \\ &= \mathbf{x}^{(k)} - 2\alpha\mathbf{A}^\top(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}) \end{aligned}$$

Supplemental material

For those who want a quick reference to matrix properties: The Matrix Cookbook

2. Newton's method is a second-order method in the setting where we consider the unconstrained, smooth convex optimization problem

$$\min_x f(x)$$

where f is convex, twice differentiable and $dom(f) = \Re^n$.
Newton's method: choose initial $x^{(0)} \in \Re^n$, and

$$x^{(k)} = x^{(k-1)} - \left(\nabla^2 f(x^{(k-1)})\right)^{-1}\nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Newton's method can be interpreted as minimizing a quadratic approximation to a function at a given point. The step $x^+ = x - \left(\nabla^2 f(x)\right)^{-1}\nabla f(x)$ can be obtained by minimizing over y the following quadratic approximation:

$$f(y) \approx f(x) + \nabla f(x)^\top(y - x) + \frac{1}{2}(y - x)^\top\nabla^2 f(x)(y - x)$$

For a quadratic one step of Newton's method minimizes the function directly because the quadratic approximation to the quadratic function will be the function itself.

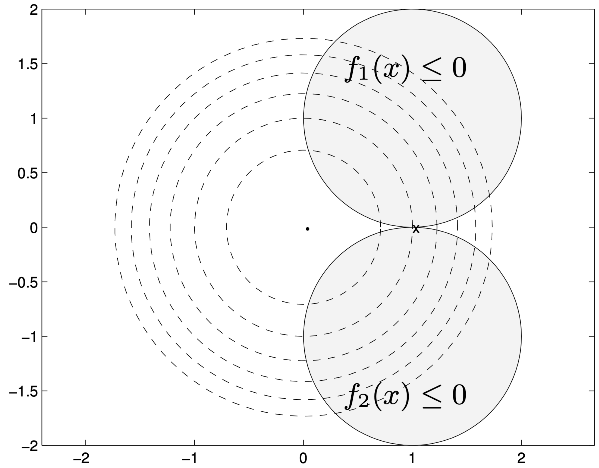
3. For $k = 0$, we get the starting point $x^{(0)} = \begin{bmatrix}\gamma \\ 1\end{bmatrix}$.

The gradient at $x^{(k)}$ is $\begin{bmatrix}x_1^{(k)} \\ \gamma x_2^{(k)}\end{bmatrix}$, so we get

$$x^{(k+1)} = x^{(k)} - \alpha\nabla f(x^{(k)}) = \begin{bmatrix}(1-\alpha)x_1^{(k)} \\ (1-\gamma\alpha)x_2^{(k)}\end{bmatrix}$$

$$\Longrightarrow x^{(k)} = \begin{bmatrix}(1-\alpha)^k x_1^{(0)} \\ (1-\gamma\alpha)^k x_2^{(0)}\end{bmatrix} = \begin{bmatrix}(1-\alpha)^k\gamma \\ (1-\gamma\alpha)^k\end{bmatrix}$$

4. (a) The figure shows the feasible set (the intersection of the two shaded disks) and some contour lines of the objective function. There is only one feasible point, (1, 0), so it is optimal for the primal problem, and we have $p^\star = 1$.



(b) The KKT conditions are

$$(x_1 - 1)^2 + (x_2 - 1)^2 \leq 1, (x_1 - 1)^2 + (x_2 + 1)^2 \leq 1,$$

$$\lambda_1 \geq 0, \lambda_2 \geq 0$$

$$2x_1 + 2\lambda_1(x_1 - 1) + 2\lambda_2(x_1 - 1) = 0$$

$$2x_2 + 2\lambda_1(x_2 - 1) + 2\lambda_2(x_2 + 1) = 0$$

$$\lambda_1((x_1 - 1)^2 + (x_2 - 1)^2 - 1) = \lambda_2((x_1 - 1)^2 + (x_2 + 1)^2 - 1) = 0.$$

At $x = (1, 0)$, these conditions reduce to

$$\lambda_1 \geq 0, \qquad \lambda_2 \geq 0, \qquad 2 = 0, \qquad -2\lambda_1 + 2\lambda_2 = 0,$$

which (clearly, in view of the third equation) have no solution.
The Lagrangian is

$$\begin{aligned}
&L(x_1, x_2, \lambda_1, \lambda_2) \\
=&x_1^2 + x_2^2 + \lambda_1((x_1 - 1)^2 + (x_2 - 1)^2 - 1) + \lambda_2((x_1 - 1)^2 + (x_2 + 1)^2 - 1) \\
=&(1 + \lambda_1 + \lambda_2)x_1^2 + (1 + \lambda_1 + \lambda_2)x_2^2 - 2(\lambda_1 + \lambda_2)x_1 - 2(\lambda_1 - \lambda_2)x_2 + \lambda_1 + \lambda_2
\end{aligned}$$

5. In publicly available solution manual.

More details

- We get equation (161) by ignoring the denominator in equation (160), since the fraction equals to 0.
- In equation (161), both a and b are scalars, thus we can switch the order of multiplication.

6. N: Normal, S: Spam

(a)

$$
\begin{aligned}
\mathrm{P}(\text{Dear} \mid \text{N}) &= 8/17; \quad \mathrm{P}(\text{Dear} \mid \text{S}) = 2/7 \\
\mathrm{P}(\text{Friend} \mid \text{N}) &= 5/17; \quad \mathrm{P}(\text{Friend} \mid \text{S}) = 1/7 \\
\mathrm{P}(\text{Lunch} \mid \text{N}) &= 3/17; \quad \mathrm{P}(\text{Lunch} \mid \text{S}) = 0 \\
\mathrm{P}(\text{Monely} \mid \text{N}) &= 1/17; \quad \mathrm{P}(\text{Monely} \mid \text{S}) = 4/7
\end{aligned}
$$

For a new message:

$$\mathrm{P}(\text{N}) = 2/3; \qquad \mathrm{P}(\text{S}) = 1/3$$

$$\mathrm{P}(\text{Dear} \mid \text{N}) \times \mathrm{P}(\text{Friend} \mid \text{N}) \times \mathrm{P}(\text{N}) \approx 0.09$$

$$\mathrm{P}(\text{Dear} \mid \text{S}) \times \mathrm{P}(\text{Friend} \mid \text{S}) \times \mathrm{P}(\text{S}) \approx 0.01$$

Hence the message is normal.

(b) Regardless of how the words are ordered, we get the same result.
Naive Bayes assumption: the features are conditionally independent given the class label.

(c) $\mathrm{P}(\text{Lunch} \mid \text{S}) = 0$. Therefore, any message containing “Lunch” has zero probability of being spam.

More details in StatQuest video.