6.7900 - Assignment 0

1.1.1: One such vector is $\frac{\mathbf{w}}{||\mathbf{w}||}$, where $\mathbf{w} = (w_1, \dots, w_n)$, since $w_0 + \sum_{i=1}^{n} w_i x_i = \mathbf{w} \cdot [(x_1, \dots, x_n) - (0, \dots, 0, \frac{w_0}{w_n})] = 0$. The equation is then $\mathbf{y} = \mathbf{v} + t\mathbf{w}$ for $\mathbf{y} \in \mathbb{R}^n$ and $t \in \mathbb{R}$.

1.1.2: We can plug the point into the function $f(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i$ and check the sign of the value.

1.1.3: The point on the hyperplane closest to $\mathbf{v}$ must satisfy $f(\mathbf{v}+t\mathbf{w}) = 0$ for some t, yielding $t = \frac{-w_0 - \mathbf{v}^T\mathbf{w}}{||\mathbf{w}||^2}$. The distance is then $||\mathbf{v} + t\mathbf{w} - \mathbf{v}|| = ||t\mathbf{w}|| = \frac{|w_0 + \mathbf{v}^T\mathbf{w}|}{||\mathbf{w}||}$.

1.2.1: Since Σ is symmetric, $\Sigma = U\Lambda U^T$ for some U and Λ such that $U = \begin{bmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_n\end{bmatrix}$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, and $\Sigma\mathbf{u}_i = \lambda_i \mathbf{u}_i$ for all i. Define $\mathbf{y} = U\mathbf{x}$ and note that $\det \frac{d\mathbf{x}}{d\mathbf{y}} = \det U^T = \det U = \pm 1$ since U is an orthogonal matrix. Then,

$$\begin{aligned}
\int_{\mathbb{R}^n} \exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x})d\mathbf{x} &= \int_{\mathbb{R}^n} \exp(-\frac{1}{2}\mathbf{y}^T\Lambda^{-1}\mathbf{y})|\det\frac{d\mathbf{x}}{d\mathbf{y}}|d\mathbf{y} \text{ by a change of variables} \\
&= \int_{\mathbb{R}^n} \exp(-\frac{1}{2}\sum_i \frac{y_i^2}{\lambda_i})d\mathbf{y} = \int_{\mathbb{R}^n} \prod_i \exp(-\frac{y_i^2}{2\lambda_i})d\mathbf{y} \\
&= \prod_i \int_{\mathbb{R}} \exp(-\frac{y_i^2}{2\lambda_i})dy_i = \prod_i \sqrt{2\pi\lambda_i} \text{ by the general Gaussian integral} \\
&= \sqrt{(2\pi)^n|\Sigma|}
\end{aligned}$$

since the product of the eigenvalues is the determinant. Thus, $\int_{\mathbb{R}^n} p_X(\mathbf{x})d\mathbf{x} = 1$ and $\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}$ is the normalizing constant. Note that Σ is assumed to be positive definite such that $\lambda_i > 0$ for all i and $\boldsymbol{\mu}$ can be assumed to be 0 since it is a constant vector.

1.2.2: By a change of variables, $p_Y(\mathbf{y}) = p_X(\frac{1}{2}\mathbf{y})\frac{dX}{dY} = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp[-2(\frac{1}{2}\mathbf{y} - \boldsymbol{\mu})^T\Sigma^{-1}(\frac{1}{2}\mathbf{y} - \boldsymbol{\mu})]$ where $\text{Cov}(Y) = 4\Sigma$.

1.2.3: The distribution has unit variance for every variable and hence is isotropic. Moreover, the variables are independent of each other since zero covariance implies independence for jointly normal variables. This can be shown by factorizing the joint PDF into a product of the marginal PDFs.

1.2.4: The distribution has variance 10 along the direction of X_1 and variance 1 along the direction of X_2. In addition, X_1 and X_2 are independent and the joint PDF can be similarly factorized into a product. The contours are ellipses with the major axes oriented along the direction of X_1.

1.2.5: The distribution has variance 10 along the directions of X_1 and X_2. In addition, X_1 and X_2 have negative covariance, indicating an inverse relationship. The contours are thin ellipses with the major axes oriented along the line $X_2 = -X_1$.

1.2.6: No, since it is not positive definite. This can be discerned by obtaining the eigenvalues 12 and -8 or noticing that $\mathbf{x}^T \begin{bmatrix} 2 & 10 \\ 10 & 2 \end{bmatrix} \mathbf{x} < 0$ for some $\mathbf{x} \in \mathbb{R}^2$.

1.2.7: This can be done by calculating $\frac{p_{X_1,X_2}(x_1,3)}{p_{X_1}(x_1)}$ or by using the conditional identities $\mathbb{E}(X_1|X_2 = 3) = \mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(3 - \mu_2)$ and $\text{Var}(X_1|X_2 = 3) = \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}$, both of which give $p_{X_1|X_2}(x_1|3) = p_{X_1}(x_1) = \frac{1}{\sqrt{20\pi}}\exp[-\frac{1}{20}(x-1)^2]$. Note that X_1 and X_2 are independent since they are jointly normal and $\text{Cov}(X_1, X_2) = 0$, as noted previously.

1.3.1: $\text{Cov}(Ax, Bx) = \mathbb{E}[(Ax - A\mathbb{E}[x])(Bx - B\mathbb{E}[x])^T] = \mathbb{E}[A(x - \mathbb{E}[x])(x - \mathbb{E}[x]^T)B^T] = A\mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x]^T)]B^T = A\text{Cov}(x)B^T$ by the linearity of expectation.

1.3.2: Denote D and ND as having and not having the disease respectively. Using Bayes' theorem, $p(D|+) = \frac{p(+|D)p(D)}{p(+)} = \frac{p(+|D)p(D)}{p(+|D)p(D)+p(+|ND)p(ND)} = \frac{(0.97)(0.00005)}{(0.97)(0.00005)+(0.03)(0.99995)} = 0.1614\%$.

1.3.3: Plot C represents $p(X)$ since there are two peaks at μ_0 and μ_1, and $P(Z = 0) = 0.8 \implies$ the density at μ_0 is greater.

1.3.4: The DAG for this situation is $A \leftarrow T \rightarrow B$, showing that A and B are independent when given T and dependent otherwise.

1.4.1: We have $\frac{\partial f}{\partial x} = 2x - y + 1$, $\frac{\partial f}{\partial y} = 4y - x - 4$, $\frac{\partial^2 f}{\partial x^2} = 2$, $\frac{\partial^2 f}{\partial y^2} = 4$, and $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} = -1$. Setting the first two to 0 yields the critical point of $(0, 1)$, which is a local minimum by the second partial derivative test and hence a global minimum. The function's value at this point is -2.

1.4.2: The Hessian $\begin{bmatrix} 2 & -1 \\ -1 & 8 \end{bmatrix}$ has eigenvalues $3 \pm \sqrt{2} > 0$, showing that it is PSD and hence convex over $\mathbb{R}^2$.

1.5.1: $\frac{\partial E}{\partial w_j} = \Sigma_i[\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) - y^{(i)}]\sigma(\mathbf{w}^T\mathbf{x}^{(i)})[1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})]x_j^{(i)} + \beta w_j$ by the chain rule and since $\sigma'(x) = \sigma(x)[1 - \sigma(x)]$.

1.5.2: The update is $w_j \leftarrow w_j - \alpha\frac{\partial E}{\partial w_j} = (1 - \beta)w_j - \alpha[\sigma(\mathbf{w}^T\mathbf{x}) - y]\sigma(\mathbf{w}^T\mathbf{x})[1 - \sigma(\mathbf{w}^T\mathbf{x})]x_j$.

1.5.3: This is not true since using only one data point in SGD introduces high variance in the update step such that individual steps might increase the objective despite decreasing it on average.

1.5.4: This is true since SGD is computationally more efficient than full gradient descent and can converge faster, among other benefits.