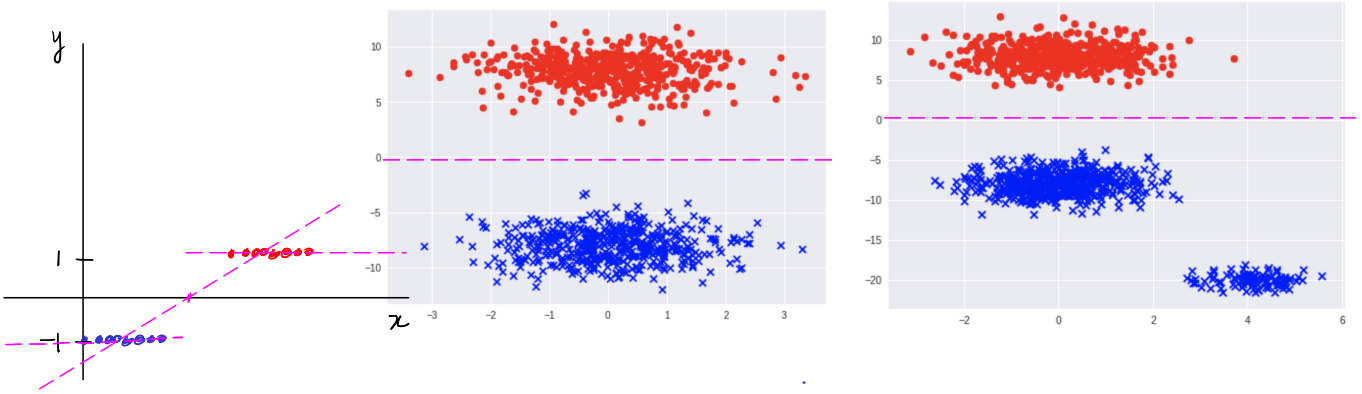
CS 189 Introduction to Machine Learning
Spring 2021 Jonathan Shewchuk

DIS7

1 Logistic Regression

In this problem, we will explore logistic regression and derive some insights.

In parts (a)–(b), we will motivate the need for logistic regression. Assume you are given the following datasets, where the red circles are one class (with label -1), and the blue X's are the other class (with label $+1$).



(a) First, suppose we are using *least-squares linear regression* to find a decision boundary that separates the two classes, where $\text{sign}(\mathbf{w}^T\mathbf{x})$ represents the classifier function. (Note that linear regression is not actually a good classification method, but we're going to briefly consider it anyway.) Draw an the decision boundary for the datasets above. Recall that the optimization problem has the form

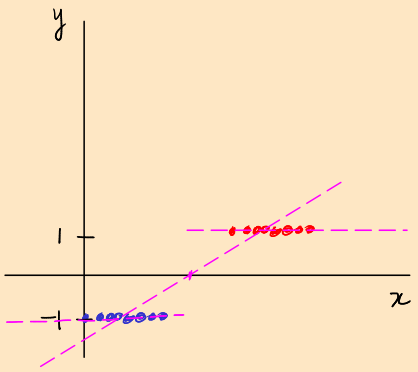
$$\arg\min_{\mathbf{w}} \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda \|\mathbf{w}\|_2^2.$$

(b) Draw the ideal decision boundary for the dataset above.

In parts (c)–(e), we will show mathematically how logistic regression tackles the issues present in least-squares linear regression, as seen above. Specifically, we will show that in logistic regression the decision boundary is less likely to be influenced by outliers of the dataset.

(c) Assume your data comes from two classes and the prior for class k is $p(y = k) = \pi_k$. Also the conditional probability distribution for each class k is Gaussian, $\mathbf{x}|(y = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, that is $f_k(\mathbf{x}) = f(\mathbf{x}|y = k) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}} \exp\Big((\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\Big)$. Assume that $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\Sigma}$ were estimated from the training data.

Show that $P(y|\mathbf{x}) = s(\mathbf{w}^\top \mathbf{x})$ is the sigmoid function, where $s(\gamma) = \frac{1}{1+e^{-\gamma}}$.



$$(c)\quad p(y=1|x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} = \left(\frac{\pi_0 f_0(x) + \pi_1 f_1(x)}{\pi_1 f_1(x)}\right)^{-1}$$

$$= \left(1 + \frac{\pi_0 f_0(x)}{\pi_1 f_1(x)}\right)^{-1}$$

$$= \left(1 + \frac{\exp(\theta_0(x))}{\exp(\theta_1(x))}\right)^{-1}$$

$$= \frac{1}{1 + \exp(-(\theta_1(x) - \theta_0(x)))}$$

$$= s\left(\theta_1(x) - \theta_0(x)\right)$$

$$w^T x$$

$$\frac{(\sqrt{2\pi})^k \pi_k}{(\sqrt{2\pi})^k |\Sigma|^{1/2}}$$

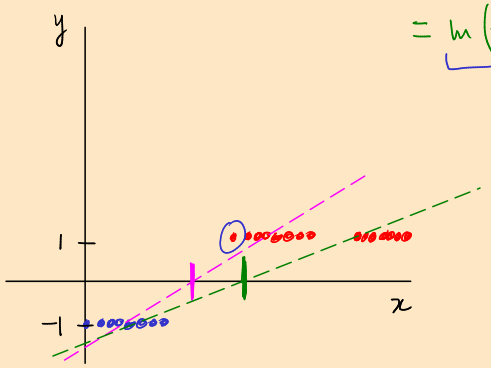
$$\theta_1(x) - \theta_0(x)$$

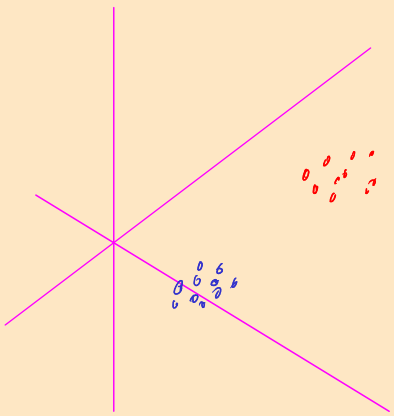
$$= \ln\left((\sqrt{2\pi})^k \pi_1 f_1(x)\right) - \ln\left((\sqrt{2\pi})^k \pi_0 f_0(x)\right)$$

$$= \ln\left(\frac{\pi_1}{|\Sigma|^{1/2}} \exp\left((x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right)\right)$$

$$- \ln\left(\frac{\pi_0}{|\Sigma|^{1/2}} \exp\left((x-\mu_0)^T \Sigma^{-1}(x-\mu_0)\right)\right)$$

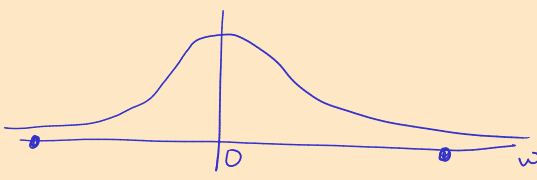
$$= \underbrace{\ln\left(\frac{\pi_1}{1-\pi_1}\right) + \frac{1}{2}\mu_0^T \Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1}_{w_0} + \underbrace{(\mu_1 - \mu_0)^T \Sigma^{-1}}_{w^T} x$$





$$c + w_1 x_1 + w_2 x_2 = 0$$

$$\Rightarrow \quad x_2 = \frac{-w_1 x_1 - c}{w_2}$$



$$s(w^T x)\left(1 - s(w^T x)\right)$$

$X^T \Omega X \Delta_{k+1} = X^T(y - s)$ $\quad w_k + \Delta_{k+1} = w_{k+1}$ $\quad \Omega = \begin{bmatrix} s_1(1-s_1) & & 0 \\ & s_2(1-s_2) & \\ 0 & & \ddots \\ & & & s_n(1-s_n) \end{bmatrix}$

$\Delta_{k+1} = (X^T \Omega X)^{-1} X^T (y - s)$

(d) In the previous part we saw that the posterior probability for each class is the sigmoid function under the LDA model assumptions. Notice that LDA is a generative model. In this part we are going to look at the discriminative model. We assume that the posterior probability has Bernoulli distribution and the probability for each class is the sigmoid function, i.e., $p(Y = y|X = \mathbf{x}; \mathbf{w}) = q^y(1-q)^{1-y}$, where $q = s(\mathbf{w}^\top \mathbf{x})$, and try to find $\mathbf{w}$ that maximizes the likelihood function. Can you find a closed form[1] maximum-likelihood estimation of $\mathbf{w}$?

(e) In this section we use Newton's method to find the optimal solution for $\mathbf{w}$. Write out the update step of Newton method. What does this say about how logistic regression handles outliers?

(d) $\mathcal{L}(w; x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n) = \prod_{i=1}^{n} p(y = y_i | x_i) = \prod_{i=1}^{n} q_i^{y_i} (1-q_i)^{1-y_i}$

$q_i := s(w^T x_i) = \frac{1}{1 + e^{-w^T x_i}}$

$\ell(w) = \log \mathcal{L}(w) = \sum_{i=1}^{n} \left[y_i \log q_i + (1-y_i) \log(1-q_i) \right] = \sum_{i=1}^{n} \left[y_i \log\left(\frac{q_i}{1-q_i}\right) + \log(1-q_i) \right]$

$1 - q_i = \frac{e^{-w^T x_i}}{1 + e^{-w^T x_i}}$

$= \sum_{i=1}^{n} y_i w^T x_i + \log\left(1 + \exp(w^T x_i)\right)$

$\log \frac{q_i}{1-q_i} = \frac{1}{1 + e^{-w^T x_i}} \cdot \frac{1 + e^{-w^T x_i}}{e^{-w^T x_i}}$

$= \frac{1}{e^{-w^T x_i}}$

$= e^{w^T x_i}$

$= w^T x_i$

$\operatorname{argmax}_w \ell(w) = \operatorname{argmin}_w -\ell(w)$

$= \operatorname{argmin}_w \underbrace{-\sum_{i=1}^{n} y_i w^T x_i - \log\left(1 + \exp(w^T x_i)\right)}_{J(w)}$

$\nabla_w J = -\sum_{i=1}^{n} y_i x_i - \frac{\exp(w^T x_i)}{1 + \exp(w^T x_i)} = X^T(s - y) = 0$ $\quad s_i = s(w^T x_i)$

$\frac{\partial}{\partial w} \log\left(1 + \exp(w^T x_i)\right)$

$= \left(\frac{\partial \log a}{\partial a}\right)\left(\frac{\partial (1 + \exp(b))}{\partial b}\right)\left(\frac{\partial}{\partial w} x_i^T w\right)$

$= \frac{1}{a} \exp(b)\, x_i^T$

$= \frac{\exp(w^T x_i)}{1 + \exp(w^T x_i)} x_i^T$

$\nabla_w \log\left(1 + \exp(w^T x_i)\right) = \left(\frac{\exp(w^T x_i)}{1 + \exp(w^T x_i)} x_i^T\right)^T = \frac{\exp(w^T x_i)}{1 + \exp(w^T x_i)} x_i$

$w_t \quad J(\cdot)$

$\nabla_w J(w_t)$

$\nabla_w^2 J(w_t)$

$\nabla_w^2 J(w_t)\, \Delta_{t+1} = -\nabla_w J(w_t)$

$w_{t+1} = \Delta_{t+1} + w_t$

$\{ x : w^T x = 0 \}$

[1] For the purpose of this question, we define a **closed form estimation** of $\mathbf{w}$ to mean an equality $\mathbf{w} = f(\mathbf{X}, \mathbf{y})$ where f is not an infinite series.

2 Bias and Variance

Oftentimes, such as in linear regression, we model the data-generating process as a noisy measurement of a true underlying response,

$$y_i = g(x_i) + \epsilon_i,$$

where ϵ_i is a zero-mean random noise variable.

We use machine learning techniques to build a hypothesis model $h(x)$ which is fit to the data as an approximation of $g(x)$. We usually don't know $g(x)$, but in the experiment that generated the plots on the next pages, suppose we know $g(x)$ is a straight line,

$$g(x) = wx + b.$$

The figures on the next pages show attempts to fit 0-degree, 1-degree, and 2-degree polynomials to g using different subsets of training data.

(a) The third figure is an attempt to fit a quadratic $h(x) = ax^2 + bx + c$ when the underlying f is a line. Why does the quadratic model learn a non-zero a? Why didn't it learn straight lines?

(b) When evaluating models, what do we mean by "bias" of a model-estimation method? Explain the differences we see in the bias for polynomials of degree 0, 1, and 2.

(c) When evaluating models, what do we mean by "variance" of a model-estimation method? Explain the differences we see in the variance for polynomials of degrees 0, 1, and 2.

(d) We can decompose the least squares risk function into bias and variance as shown in lecture.

$$\begin{aligned}
\mathbb{E}[(h(x) - y)^2] &= \mathbb{E}[h(x)^2] + \mathbb{E}[y^2] - 2\mathbb{E}[y(h(x)] \\
&= \text{Var}(h(x)) + \mathbb{E}[h(x)]^2 + \text{Var}(y) + \mathbb{E}[y]^2 - 2\mathbb{E}[y]\mathbb{E}[h(x)] \\
&= (\mathbb{E}[h(x)] - \mathbb{E}[y])^2 + \text{Var}(h(x)) + \text{Var}(y) \\
&= \underbrace{(\mathbb{E}[h(x) - g(x)])^2}_{\text{bias squared of method}} + \underbrace{\text{Var}(h(x))}_{\text{variance of method}} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible error}}
\end{aligned}$$

We can decompose the error this way over the entire dataset, or we can decompose an individual point's error into these three components.

Now, observe the last figure. Why is the variance larger for points near the left and right extremes, and smaller for points in the middle?

(e) Why is our estimate of the bias not zero for the degree-1 and degree-2 models? Would it be zero if we generated an infinite number of datasets?

(f) How are bias and variance related to overfitting and underfitting?

(g) Does training error provide a measure of bias, variance, or both? How about validation and test error?

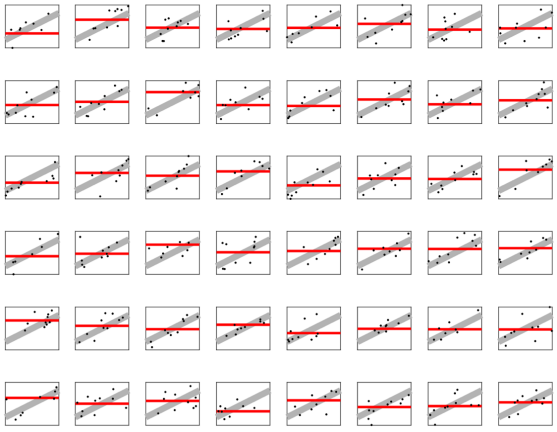
(h) How can we interpret the bias-variance trade-off in hard- and soft-margin SVM? Recall that the soft margin SVM objective is

$$\min \|w\|^2 + C \sum_i \xi_i \quad \text{subject to} \quad y_i(x_i^\top w + \alpha) \geq 1 - \xi_i; \quad \xi_i \geq 0.$$

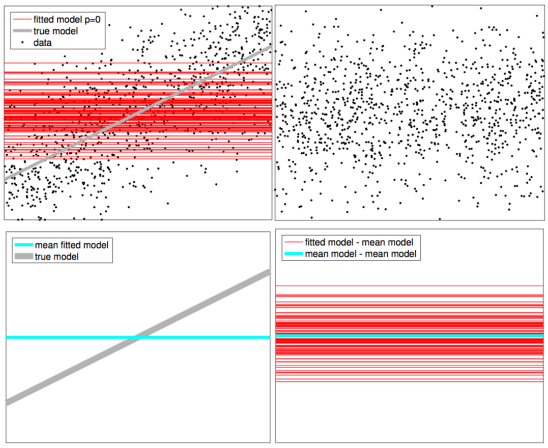
(i) How can we interpret the bias-variance trade-off in LDA and QDA?

The figures on the left show many different models fit on subsets of training data for degrees $p = 0, 1, 2$. The figures on the right, the top left shows all learned models on top of the true model and data. The top right shows the noise of each data point, or the residual after subtracting $y - g(x)$. The bottom left shows the average learned model on top of the true model, and the figure on the bottom right shows all learned models on top of the average learned model.

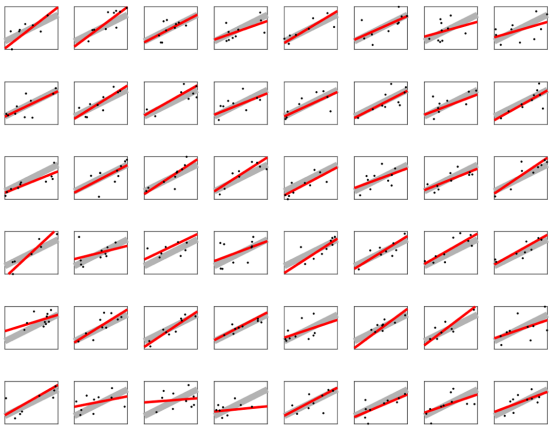
Fitting A Model over Multiple Datasets: $p = 0$



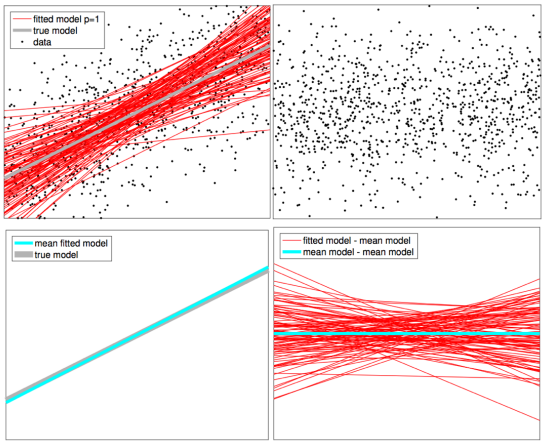
Bias and Variance in Model Selection: $p = 0$



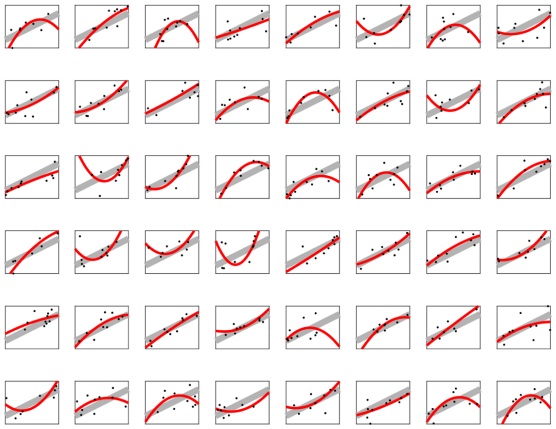
Fitting A Model over Multiple Datasets: $p = 1$



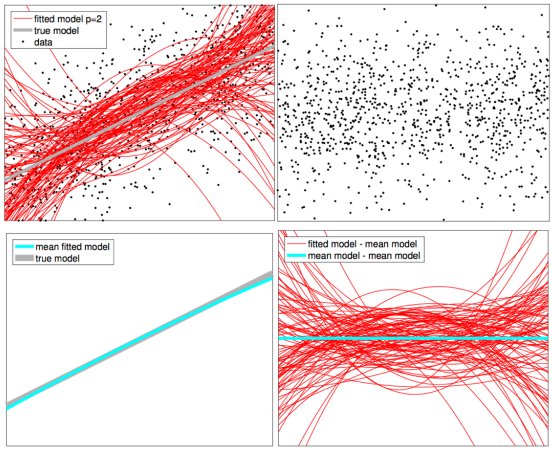
Bias and Variance in Model Selection: $p = 1$



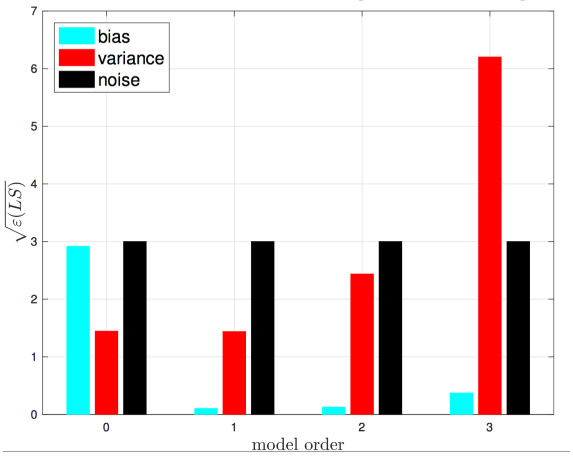
Fitting A Model over Multiple Datasets: $p = 2$



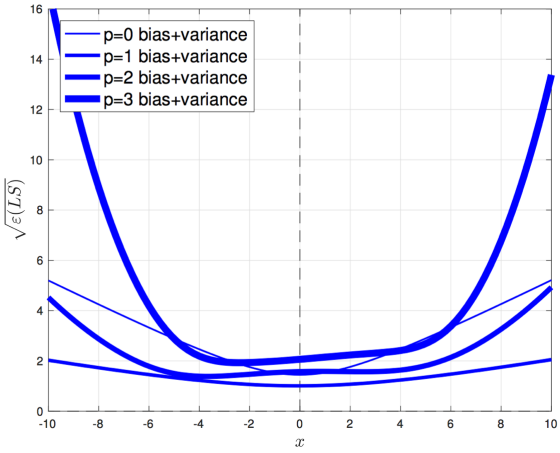
Bias and Variance in Model Selection: $p = 2$



Bias and Variance: Underfitting vs. Overfitting



Variation of Prediction Error with Model Order



3 Logistic Posterior with Different Variances

In Discussion 3, we proved that under an exponential class-conditional distribution, the posterior could be written in the form of a sigmoid that was linear over x. In this problem, we show that the posterior of a univariate QDA problem can also be written in the form of a sigmoid, but now it is *quadratic* over x. Consider the case when the class conditionals are Gaussian, but have different variances, i.e.,

$$(X|Y=i) \sim \mathcal{N}(\mu_i, \sigma_i^2), \quad \text{where } i \in \{0,1\}$$
$$Y \sim \text{Bernoulli}(\pi)$$

1. Show that the posterior distribution of the class label given X is also a logistic function, but with a quadratic argument in X. That is, show that $P(Y=1|X=x)$ is of the form $1/(1+e^{-h(x)})$, where $h(x) = ax^2 + bx + c$ is quadratic in x.
2. Assuming 0-1 loss, what will the decision boundary look like (i.e., describe what the posterior probability plot looks like)?
3. Now suppose that we are dealing with an asymmetric loss function. Describe how this changes the decision boundary, if at all.

1. $P(Y=1|x) = \dfrac{f(x|Y=1)\Pr[Y=1]}{f(x|Y=1)\Pr[Y=1] + f(x|Y=0)\Pr[Y=0]}$

$= \dfrac{1}{\dfrac{f(x|Y=1)\Pr[Y=1] + f(x|Y=0)\Pr[Y=0]}{f(x|Y=1)\Pr[Y=1]}}$

$= \dfrac{1}{1 + \dfrac{f(x|Y=0)\Pr[Y=0]}{f(x|Y=1)\Pr[Y=1]}}$ $\dfrac{1-\pi}{\pi}$ $= \dfrac{1}{1 + \frac{\sigma_1}{\sigma_0}\frac{1-\pi}{\pi}\exp\left(\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\right)}$

$\exp\left[\left(\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_0^2}\right)x^2 + \left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right)x + \left(\frac{\mu_1^2}{2\sigma_1^2} - \frac{\mu_0^2}{2\sigma_0^2} + \ln\left(\frac{\sigma_1}{\sigma_0}\frac{1-\pi}{\pi}\right)\right)\right]$

a b c

$h(x) = ax^2 + bx + c$

$C\exp(x) = \exp(\ln(C))\exp(x) = \exp(\ln(C) + x)$

2. $P(Y=1|x) = P(Y=0|x)$
3. Predict 1 if expected loss (risk) of pred 1 < expected loss (risk) of pred 0

$P(Y=1|x)\,L(0,1) = P(Y=0|x)\,L(1,0)$

4 Multivariate Gaussians: A Review

Consider a two dimensional random variable $Z \in \mathbb{R}^2$. In order for the random variable to be jointly Gaussian, a necessary and sufficient condition is that

- Z_1 and Z_2 are each marginally Gaussian, and
- $Z_1|Z_2 = z$ is Gaussian and $Z_2|Z_1 = z$ is Gaussian.

Recall that the PDF of a multivariate Gaussian is $f(z) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$.

(a) Let X_1 and X_2 be i.i.d. standard normal random variables. Let U be a discrete random variable such that $P(U = -1) = P(U = 1) = \frac{1}{2}$, independent of everything else. First, verify if the conditions of the first characterization hold for the following random variables (i.e., they are each marginally Gaussian, and their conditional probabilities are also Gaussian). Second, calculate the covariance matrix Σ_Z.

   (a) $Z_1 = X_1$ and $Z_2 = X_2$.
   (b) $Z_1 = X_1$ and $Z_2 = -X_1$.
   (c) $Z_1 = X_1$ and $Z_2 = UX_1$.

(b) Use the example above to show that two Gaussian random variables can be uncorrelated, but not independent. On the other hand, show that two uncorrelated, jointly Gaussian random variables are independent.

(c) Let $Z = VX$, where $V \in \mathbb{R}^{2\times 2}$, $Z, X \in \mathbb{R}^2$, and $X \sim N(0, I)$. What is the covariance matrix Σ_Z? Is this also true for a random variable other than Gaussian?

(d) Use the above setup to show that $X_1 + X_2$ and $X_1 - X_2$ are independent. Give another example pair of linear combinations that are independent.

5 Gradient Descent and Convexity

The smoothed version of the hinge loss function[2] with parameter t is

$$f(y) = \begin{cases} \frac{1}{2} - ty & \text{if } ty \le 0, \\ \frac{1}{2}(1-ty)^2 & \text{if } 0 < ty < 1, \\ 0 & \text{if } 1 \le ty. \end{cases}$$

$(-t)\left(\frac{1}{2}(1-ty)\right)$

Define $L(w) = \frac{1}{n}\sum_{i=1}^{n} f(w^\top x_i - y_i)$. Given sample points $x_1, x_2, \ldots, x_n$ and labels $y_1, y_2, \ldots, y_n$, we define the optimization problem

$$\min_w \frac{1}{n}\sum_{i=1}^{n} f(w^\top x_i - y_i)$$

1. Is $L(w)$ convex?
2. Write out the gradient descent update equation.
3. Write out the stochastic gradient descent update equation.

1. $g(y) = \frac{\partial f}{\partial y}$

$$g(y) = \begin{cases} -t & ty \le 0 \\ -t(1-ty) & 0 < ty < 1 \\ 0 & 1 \le ty \end{cases}$$

$-t + t^2 y$

$$h(y) = \frac{\partial^2 f}{\partial y^2} = \frac{\partial g}{\partial y} = \begin{cases} 0 & ty \le 0 \\ t^2 & 0 < ty < 1 \\ 0 & 1 \le ty \end{cases}$$

$h(y) \ge 0 \ \forall y \in \mathbb{R}$. $\therefore f(y)$ is convex. $\therefore L(w)$ is conv $\because$ it is a sum of convex funcs.

2. $\frac{\partial f(w^\top x - y)}{\partial w} = g(w^\top x - y)\frac{\partial (w^\top x - y)}{\partial w} = g(w^\top x - y)x^\top$. $\nabla_w f(w^\top x - y) = g(w^\top x - y)x$

$\nabla_w f(w^\top x_i - y_i) = g(w^\top x_i - y_i)x_i$

$\nabla_w L(w) = \frac{1}{n}\sum_{i=1}^{n} g(w^\top x_i - y_i)x_i$

$w^{(k+1)} \leftarrow w^{(k)} - \alpha \nabla_w L(w^{(k)}) = w^{(k)} - \alpha \frac{1}{n}\sum_{i=1}^{n} g(w^{(k)\top} x_i - y_i)x_i$

3. $w^{(k+1)} \leftarrow w^{(k)} - \alpha \nabla_w L(w^{(k)}) = w^{(k)} - \alpha g(w^{(k)\top} x_i - y_i)x_i$

[2] This function is not required knowledge.