# Machine Learning for Finance – Problem Set 4

Neal Parikh

March 13, 2018

*Instructions.* Do not refer to any outside sources to complete this assignment, in accordance with the honor code. If you discussed any problems with other students, indicate that in your solutions.

1. *Gaussian discriminant analysis.* Consider a dataset $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i \in \mathbf{R}^n$ and $y_i \in \{0, 1\}$, and consider the following model for the joint distribution $p(x, y)$:

$$\begin{array}{rcl} y & \sim & \mathrm{Bernoulli}(\phi) \\ x \,|\, y = 0 & \sim & \mathrm{N}(\mu_0, \Sigma) \\ x \,|\, y = 1 & \sim & \mathrm{N}(\mu_1, \Sigma), \end{array}$$

with parameters $\phi$, $\mu_0$, $\mu_1$, and $\Sigma$.

(a) Suppose we already have estimates of all the four parameters and now want to make a prediction at a new query point $x^{\text{new}}$. Show that the posterior distribution of the label at $x^{\text{new}}$ takes the form of a logistic function and can be written as

$$p(y = 1 \,|\, x) = \frac{1}{1 + \exp(-\theta^T x)},$$

where $\theta$ is a function of $\phi$, $\mu_0$, $\mu_1$, and $\Sigma$. (To get your answer into the form above, you may need to add a constant feature 1 into $x_i$ and consider them as vectors in $\mathbf{R}^{n+1}$.) This implies, for instance, that linear discriminant analysis is a linear classifier.

(b) Show that the maximum likelihood estimates of the model parameters are given by the following expressions:

$$\begin{array}{rcl} \hat{\phi} & = & \displaystyle\frac{1}{N}\sum_{i=1}^{N}[y_i = 1], \\ \hat{\mu}_k & = & \displaystyle\frac{\sum_{i=1}^{N}[y_i = k]x_i}{\sum_{i=1}^{N}[y_i = k]}, \\ \hat{\Sigma} & = & \displaystyle\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{y_i})(x_i - \mu_{y_i})^T. \end{array}$$

*Note.* You can use that $\nabla f(X) = X^{-1}$ when $f(X) = \log\det X$ with $\mathbf{dom}\, f = \mathbf{S}^n_{++}$, and that $\nabla f(X) = aa^T$ when $f(X) = a^T X a$.

2. Consider fitting an SVM to a linearly separable training set. Is the SVM guaranteed to choose a decision boundary that separates the positive and negative classes?

3. *Logistic regression and SVMs as regularized loss minimization.* It turns out that (regularized) logistic regression and the support vector machine optimize related objective functions.

   (a) Given an example $x \in \mathbf{R}^n$ and a label $y \in \{0, 1\}$, the log likelihood of the example under the logistic regression model is

   $$y \log g(x) + (1 - y) \log(1 - g(x)),$$

   where $g(z) = 1/(1 + \exp(-z))$ is the sigmoid function. Now suppose we want to switch to the convention of using the labels $\tilde{y} \in \{-1, 1\}$. Show that

   $$y \log g(x) + (1 - y) \log(1 - g(x)) = -\log(1 + \exp(-\tilde{y} w^T x)).$$

   In other words, the expression on the righthand side gives the likelihood of a single example when using the $\{-1, 1\}$ label convention.

   (b) Given a training set $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, with $y_i \in \{-1, 1\}$. Show that the maximum likelihood estimate of the parameters is given by minimizing

   $$\sum_{i=1}^{N} \log(1 + \exp(-y_i w^T x_i)).$$

   (c) Consider the function $f(z) = \log(1 + \exp(-z))$. Explain in English what happens to $f$ under the limits $z \to \infty$ and $z \to -\infty$. Sketch or plot the shape of $f$ and indicate asymptotes and intercepts in your sketch.

   (d) Show that the problem

   $$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & y(w^T x + b) \geq 1 - t \\ & t \geq 0 \end{array}$$

   with variables $w \in \mathbf{R}^n$, $b \in \mathbf{R}$, and $t \in \mathbf{R}$ is equivalent to the unconstrained problem

   $$\text{minimize} \quad (1 - y(w^T x + b))_+$$

   with variables $w$ and $b$.

   (e) Draw the function $f(z) = (1 - z)_+$ and indicate asymptotes and intercepts. The function $f$ is known as *hinge loss*; explain why this makes sense.

   (f) The problem in part (d) was a version of an SVM estimation problem with a single training example. Generalize your derivation for part (d) to derive an unconstrained problem equivalent to the problem

   $$\begin{array}{ll} \text{minimize} & (1/2)\|w\|_2^2 + \lambda \mathbf{1}^T t \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 - t_i, \quad i = 1, \ldots, N \\ & t \succeq 0, \end{array}$$

   with variables $w \in \mathbf{R}^n$, $b \in \mathbf{R}$, $t \in \mathbf{R}^N$.

4. *Invariance properties of support vector machines.* Recall that training an SVM involves solving the convex optimization problem

$$\begin{array}{ll} \text{minimize} & (1/2)\|w\|_2^2 + \lambda \mathbf{1}^T t \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 - t_i, \quad i = 1, \ldots, N \\ & t \succeq 0, \end{array} \tag{1}$$

with variables $w \in \mathbf{R}^n$, $b \in \mathbf{R}$, $t \in \mathbf{R}^N$.

Suppose $\theta^\star = (w^\star, b^\star, t^\star)$ is the solution to the problem above. Here, we look at some ways in which the optimization problem we solve to fit an SVM model can be changed without changing the predictions made by the resulting classifier.

(a) *Margin scaling.* One of the decisions that may have seemed arbitrary but possibly significant in the formulation of the SVM was the choice of scaling the margin boundaries to lie at the hyperplanes $w^T x + b = 1$ and $w^T x + b = -1$. Suppose we replace the 1 in the margin constraint of the SVM with some constant $\kappa > 0$ and replace the regularization parameter $\lambda$ with $\kappa\lambda$, giving the problem

$$\begin{array}{ll} \text{minimize} & (1/2)\|w\|_2^2 + \kappa\lambda \mathbf{1}^T t \\ \text{subject to} & y_i(w^T x_i + b) \geq \kappa - t_i, \quad i = 1, \ldots, N \\ & t \succeq 0. \end{array} \tag{2}$$

Let $\tilde{\theta}^\star = (\tilde{w}^\star, \tilde{b}^\star, \tilde{t}^\star) = \kappa(w^\star, b^\star, t^\star)$.

i. Show that $\tilde{\theta}^\star$ is feasible for (2).

ii. Show that $\tilde{\theta}^\star$ is the solution of (2).

iii. Show that the modified SVM makes the same classification decisions, *i.e.*, that $\mathbf{sign}((\tilde{w}^\star)^T x + \tilde{b}^\star) = \mathbf{sign}((w^\star)^T x + b^\star)$.

(b) *Orthogonal invariance of Gaussian kernels.* Particular forms of the SVM, such as the SVM with some choice of kernel, may have additional invariances. Recall the kernelized dual SVM problem

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T \alpha - (1/2) \sum_{i,j=1}^N y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\ \text{subject to} & \alpha^T y = 0 \\ & 0 \preceq \alpha \preceq \lambda \mathbf{1}, \end{array} \tag{3}$$

with variable $\alpha \in \mathbf{R}^N$. Suppose $K$ is chosen to be the Gaussian kernel

$$K(x, z) = \exp\left(-\frac{\|x - z\|_2^2}{2\sigma^2}\right).$$

Let $\mathcal{D} = \{(x_i, y_i) \mid i = 1, \ldots, N\}$ be a training set and let $\tilde{\mathcal{D}} = \{(Qx_i, y_i) \mid i = 1, \ldots, N\}$ be a modified dataset where the input vectors $x_i$ have been transformed by an orthogonal matrix $Q \in \mathbf{R}^{n \times n}$.

i. Let $\alpha^\star$ and $\tilde{\alpha}^\star$ be the solution to the dual problem when using training sets $\mathcal{D}$ and $\tilde{\mathcal{D}}$, respectively. Show that $\alpha^\star = \tilde{\alpha}^\star$.

ii. Let $\hat{f}(x)$ and $\tilde{f}(x)$ denote predictions on example $x$ of SVMs trained on $\mathcal{D}$ and $\tilde{\mathcal{D}}$, respectively. Which of the following are true?

(A) $\hat{f}(x) = \tilde{f}(x)$

(B) $\hat{f}(x) = \tilde{f}(Qx)$

(C) $\hat{f}(Qx) = \tilde{f}(x)$

You can use the following expression for the optimal value of $b$ for $\tilde{f}$:

$$b = -\frac{\max_i [y_i = -1] w^T \varphi(Qx_i) + \min_i [y_i = 1] w^T \varphi(Qx_i)}{2},$$

where $\varphi$ is the feature map associated with the Gaussian kernel.

*Note.* The output of $\varphi$ for a Gaussian kernel is actually infinite dimensional, so the inner product notation $x^T z$ should be written differently, but this small abuse of notation does not affect the problem.

5. *A simple duality example.* Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & x^2 + 1 \\ \text{subject to} & (x-2)(x-4) \leq 0, \end{array}$$

with variable $x \in \mathbf{R}$.

(a) *Analysis of primal problem.* Give the feasible set, the optimal value, and the optimal solution.

(b) *Lagrangian and dual function.* Plot the objective $x^2 + 1$ versus $x$. On the same plot, show the feasible set, optimal point and value, and plot the Lagrangian $L(x, \lambda)$ versus $x$ for a few positive values of $\lambda$. Verify the lower bound property ($p^\star \geq \inf_x L(x, \lambda)$ for $\lambda \geq 0$). Derive and sketch the Lagrange dual function $g$.

(c) *Lagrange dual problem.* State the dual problem, and verify that it is a concave maximization problem. Find the dual optimal value and dual optimal solution $\lambda^\star$. Does strong duality hold?

6. *Kernel logistic regression.* Models other than the support vector machine can be kernelized, *i.e.*, put in a form where the kernel trick can be used. Recall that a (regularized) logistic regression model can be fit by solving the convex optimization problem

$$\text{minimize} \quad -\textstyle\sum_{i=1}^N \log p(y_i \mid x_i; w) + (\lambda/2)\|w\|_2^2 \tag{4}$$

with variable $w \in \mathbf{R}^n$, where $p(y = 1 \mid x) = s(w^T x)$ and $s(z) = 1/(1 + \exp(-z))$ is the sigmoid function. (Here, the inputs $x_i$ are assumed to already contain a constant term and so the separate intercept parameter $b$ is omitted.)

(a) Consider a modified version of logistic regression in which the labels are $\{-1, 1\}$ instead of $\{0, 1\}$, and the conditional probability of the label is given by

$$p(y = k \mid x; w) = s(yw^T x)$$

for $k \in \{-1, 1\}$. Show that this implies

$$p(y = 1 \,|\, x) = s(w^T x), \quad p(y = -1 \,|\, x) = 1 - s(w^T x).$$

(b) Show that (4) is equivalent to the problem

$$\begin{array}{ll} \text{minimize} & (\lambda/2)\|w\|_2^2 + \sum_{i=1}^N \log(1 + \exp(-u_i)) \\ \text{subject to} & u_i = y_i w^T x_i, \quad i = 1, \ldots, N, \end{array} \tag{5}$$

with variables $w \in \mathbf{R}^n$, $u \in \mathbf{R}^N$. It should be evident that (5) is convex.

(c) The Lagrangian for (5) is

$$L(w, u, \alpha) = \frac{\lambda}{2}\|w\|_2^2 + \sum_{i=1}^N \log(1 + e^{-u_i}) + \sum_{i=1}^N \alpha_i (u_i - y_i w^T x_i),$$

with dual variables $\alpha \in \mathbf{R}^N$. Find an expression for the $\hat{w}$ minimizing $L$ for a fixed value of $\alpha$; the expression will be in terms of $\alpha$. Explain why this expression implies that $0 \preceq \alpha \preceq \mathbf{1}$.

(d) Show that the dual objective is given by

$$g(\alpha) = -\sum_{i=1}^N (\alpha_i \log \alpha_i + (1 - \alpha_i)\log(1 - \alpha_i)) - \frac{1}{2\lambda}\sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j x_i^T x_j.$$

(e) Given this dual formulation of the (regularized) logistic regression problem, explain how to kernelize logistic regression. In particular, describe how to train the model and how to efficiently compute $p(y = 1 \,|\, x^{\text{new}})$ for a new test input $x^{\text{new}}$.

7. *Constructing kernels.* Kernelized algorithms let us significantly increase the power of a relatively simple method by allowing it to implicitly work in a high-dimensional space, but a main question is then how to construct kernels for a given problem.

We have seen two main ways to construct kernels: (a) explicitly define a feature map $\varphi$, and (b) use Mercer's theorem. This exercise will build on these by allowing us to construct new kernels from existing ones.

For each of the functions $K$ below, state whether or not it is a kernel. If so, prove it; otherwise, provide a counterexample.

(a) $K(x, z) = K_1(x, z) + K_2(x, z)$, where $K_1, K_2$ are kernels on $\mathbf{R}^n$.

(b) $K(x, z) = K_1(x, z) - K_2(x, z)$.

(c) $K(x, z) = \alpha K_1(x, z)$, where $\alpha > 0$.

(d) $K(x, z) = -\alpha K_1(x, z)$.

(e) $K(x, z) = K_1(x, z) K_2(x, z)$.

(f) $K(x, z) = f(x) f(z)$, where $f : \mathbf{R}^n \to \mathbf{R}$. What are the implications if $f$ is a density?

(g) $K(x, z) = K_3(T(x), T(z))$, where $T : \mathbf{R}^n \to \mathbf{R}^d$.

(h) $K(x, z) = p(K_1(x, z))$, where $p$ is a polynomial with positive coefficients.